

UNIVERSAL
LIBRARY



133 460

UNIVERSAL
LIBRARY

EDUCATIONAL MEASUREMENTS
AND THE
CLASS-ROOM TEACHER

The Century Education Series

Educational Measurements AND THE Class-Room Teacher

BY

A. R. GILLILAND

PROFESSOR OF PSYCHOLOGY, NORTHWESTERN UNIVERSITY

R. H. JORDAN

PROFESSOR OF EDUCATION, CORNELL UNIVERSITY

FRANK S. FREEMAN

ASSISTANT PROFESSOR OF EDUCATION, CORNELL UNIVERSITY



THE CENTURY CO.

New York & London

COPYRIGHT, 1931, BY THE CENTURY CO.
ALL RIGHTS RESERVED, INCLUDING THE
RIGHT TO REPRODUCE THIS BOOK, OR
OPTIONS THEREOF, IN ANY FORM. 3110

PRINTED IN U. S. A.

PREFACE

In presenting an addition to the list of books treating of the general subject of educational measurements, some definite justification must be offered. The authors of the present text feel that there is a twofold need which has been but partially met by the otherwise admirable books now before the public. The first is the need for a text which may be used as a handbook for guidance of the teacher in service; the second is the need for a class-room text adapted to the use of prospective class-room teachers.

As to the first, there seems to be a haziness in the minds of many class-room teachers in both elementary and secondary schools as to the use they may make of tests and scales in their own work. They are too often of the opinion that tests are primarily supervisory instruments, and so too difficult in operation and too abstruse in interpretation to be of any real aid to the individual teacher. A certain emphasis, then, should be put upon the fact that achievement tests are valuable instruments for the teacher to understand and use independently of, or in coöperation with, the supervisor. The teacher should understand that she may handle her own work more intelligently, with more successful results, in proportion as she makes the greater use of proper measures of her effort.

Again, there has been confusion in the terminology

of tests, which has given rise to many false notions among the rank and file of the teaching profession. The expression "psychological tests" has been used as a blanket phrase to cover every sort of measurement, and many teachers have not learned to discriminate between various types of psychological measures. Comparatively few teachers, to judge from reactions in representative summer school groups, can, for example, explain accurately the difference between intelligence tests and achievement scales. Thus an unfortunate situation holds which must be cleared up in order that the work of the clinician and supervisor may be done with assurance of understanding and resultant sympathy from the body of class-room instructors.

The second great need refers to two aspects of the work in classes in education in normal schools and colleges which deserve a rather different emphasis than has hitherto been given. On the one hand is the class of students preparing primarily for class-room teaching, who wish to study the technique and meaning of achievement tests at the same time that they are given practice in their use and evaluation. In the single semester frequently given to this work, they do not wish to include a consideration of the intelligence test or other psychological measures, and thus need a manual with the major emphasis upon the use of the achievement test. The study of the intelligence test will come in a separate course, as will the factors of supervision and administration as related to measurement.

On the other hand, we frequently meet a group within college departments of education, made up of undergraduates who are concerned more with general method

and subject matter of instruction than in the psychological phases of the teaching process. Unfortunately there are still a great many of this sort of students; and the teacher of general methods of instruction finds that they have no idea of the meaning or use of the tests, and that frequently they do not have a place in their undergraduate programs for such special courses. There is a place for a manual which may accompany such courses in general method, to give this knowledge of the various tests and their uses, and so to prevent the student from going into the work entirely ignorant of the great possibilities of such instruments. The authors have had the thought of working out a textbook which may be adapted to both types of undergraduate classes.

The present volume, then, is an attempt to meet this twofold need; in order to meet the varying objectives the idea has been kept steadily in mind of presenting the important concepts and methods of application of tests and scales in as simple and non-technical language as is consistent with sound practice in their employment. The order of topics and general arrangement have been planned from this point of view. It is hoped that the materials are so presented that the book will find its place as a manual for the teacher in service and a suitable part of the reading circle on the one hand, and as a class-room text for the normal school, college, and school of education on the other.

It is impossible to acknowledge all sources from which materials for this text have been drawn. Some materials have been taken directly; others have been included in modified form, while still other materials

have served only as suggestions for our purpose. In the subject of educational measurements the fund of materials is so great that original sources are often very difficult to discover. Yet the authors have attempted to acknowledge their indebtedness wherever possible.

In addition to acknowledgments to publishers of text-books from which material has been quoted, special acknowledgments should be made to the many publishers and authors of the tests and scales from which quotations and illustrations have been freely taken. Chief among these are: The Public School Publishing Co., World Book Co., Bureau of Publications, Teachers College, Columbia University, Russell Sage Foundation and Mr. S. A. Courtis.

The authors are especially indebted to Miss Margaret Gessford of Washington, D. C., who painstakingly read the original manuscript and gave many valuable suggestions for its revision.

THE AUTHORS.

PREFACE TO THE REVISED EDITION

In the field of educational measurements, new tests frequently appear, and the older tests are being constantly subjected to scrutiny and experiment for the purpose of evaluation. This revision, therefore, has been made in the light of the contributions during the six years since the appearance of the first edition. Where justified the less significant of the older tests have been eliminated from the text; and new tests of demonstrated merit have been included. Furthermore, older tests have been re-evaluated in the light of recent experiment. Most of the chapters have been enlarged, particularly those dealing with secondary school subjects. The discussion of statistics and graphic method has been expanded, primarily from the point of view of interpretation. In the general discussion (Chaps. I, II, III, and IV), additions and modifications were made wherever the need was indicated by more recent investigation. Throughout the text, however, we have held to the point of view as indicated in the preface to the first edition.

THE AUTHORS.

December, 1930

CONTENTS

	PAGE
CHAPTER I. REASONS FOR EDUCATIONAL MEASUREMENTS . . .	3
Traditional methods of grading—the need for more adequate methods of grading— <i>selected references</i>	
CHAPTER II. WHAT CONSTITUTES A STANDARD MEASURE . . .	20
Factors involved in standard tests—differentiations from traditional measures—necessary specifications of the objective measure—the criteria of a good test—different kinds of measures—mental tests— <i>selected references</i>	
CHAPTER III. THE PRACTICAL USE OF EDUCATIONAL MEASURES IN THE CLASS-ROOM	40
Introduction—uses of achievement tests—for purposes of comparison—for diagnosis of teaching—for evaluating methods of teaching—for diagnosis of classes—for diagnosis of individual pupils—for setting standards of achievement—for promotion—uses of intelligence tests—for class diagnosis—for individual diagnosis—the combination of achievement and intelligence test results—summary— <i>selected references</i>	
CHAPTER IV. REQUISITES FOR GIVING OBJECTIVE TESTS . . .	67
The technique of objective tests—administering the tests—scoring the papers—interpretation of the scores	
CHAPTER V. SPELLING	78
Early objective tests—some problems of spelling—Ayles Spelling Scale—Buckingham Extension of the Ayles Scale—Monroe's Timed Spelling Tests—Iowa Spelling Scale—Material of English Spelling, Jones—The Teachers Word Book, Thorndike—Morrison-McCall Spelling Scale—summary—materials needed—supplementary list of tests— <i>selected references</i>	

	PAGE
CHAPTER VI. HANDWRITING	98
Problems in the measurement of handwriting—Freeman Analytical Handwriting Scale—Ayres Measuring Scale for Handwriting—Thorndike Scale for the Handwriting of Children in Grades Five to Eight—Gray Standard Score Card for Measuring Handwriting—Courtis Standard Practice Tests in Handwriting— <i>materials needed—supplementary list of scales—selected references</i>	
CHAPTER VII. READING	123
Importance and problems of reading—types of tests—Gray Standardized Oral Reading Paragraphs and the Oral Reading Check Tests—Pressey First Grade Word Reading Test and First Grade Reading Scale—Thorndike Visual Vocabulary Scale and the Test of Word Knowledge—Monroe Standardized Silent Reading Test—Thorndike-McCall Reading Scale—Burgess Scale for Measuring Ability in Silent Reading—Courtis Silent Reading Test—Gray Silent Reading Test—Haggerity Reading Examination— <i>materials needed—supplementary list of tests—selected references</i>	
CHAPTER VIII. ENGLISH LANGUAGE AND COMPOSITION . . .	158
The problem of measuring language ability—types of language tests—Starch Punctuation Scale—Starch's English Grammar Scales—Charters' Diagnostic Language Tests—New York English Survey Tests—Cross English Test—Trabue Completion Language Scales—tests of composition—Nassau County Supplement to the Hillegas Scale—other composition scales—Abbott and Trabue Exercises in Judging Poetry—conclusions— <i>materials needed—supplementary list of tests—selected references</i>	
CHAPTER IX. ARITHMETIC	182
Importance and problems of instruction in arithmetic—types of arithmetic tests—Courtis Standard Research Tests, Series B—Courtis Standard Practice Tests in Arithmetic—Compass Diagnostic Tests in Arithmetic—Cleveland Survey Arithmetic Tests—Woody-McCall Mixed Fundamentals—Monroe Diagnostic Tests in Arithmetic—Monroe's Standardized Reasoning Tests in Arithmetic—New Stone Reasoning Test in Arithmetic—conclusions— <i>materials needed—supplementary list of tests—selected references</i>	

CONTENTS

xiii

PAGE

208

CHAPTER X. GEOGRAPHY	208
The problem of measuring in geography—Hahn-Lackey Geography Scale—Posey-Van Wagenen Geography Scales—Buckingham-Stevenson Place Geography Tests—Courtis Supervisory Tests in Geography—Branom's Diagnostic Tests in Geography—conclusions—materials needed—supplementary lists of tests—selected references	
CHAPTER XI. HISTORY AND CIVICS	222
The problem of measurement in history—Hahn History Scale—Barr Diagnostic Tests in American History—Harlan Test of Information in American History—Van Wagenen American History Scales—Pressey-Richards Tests in the Understanding of American History—Columbia Research Bureau American History Test—Brown-Woody Civics Test—conclusions—materials needed—supplementary list of tests—selected references	
CHAPTER XII. MUSIC AND DRAWING	246
Music as a special talent—Seashore Measure of Musical Talent—Courtis Standard Supervisory Tests in Music—Kwalwasser-Ruch Test of Musical Accomplishment—Thordike Scale for the Merit of Drawing by Pupils Eight to Fifteen Years Old—Kline-Carey Measuring Scale for Free-hand drawing—conclusions—materials needed—supplementary list of tests—selected references	
CHAPTER XIII. SECONDARY SCHOOL MATHEMATICS	261
The problem of measurement in mathematics—Rogers Test of Mathematical Ability—Kelley Mathematical values Test Alpha—Hotz First Year Algebra Scales—Douglass Standard Diagnostic Tests for Elementary Algebra—Illinois Standardized Algebra Tests—Minnick Geometry Tests—Sehorling-Sanford Achievement Test in Plane Geometry—Columbia Research Bureau Plane Geometry Test—conclusions—materials needed—supplementary list of tests—selected references	
CHAPTER XIV. SECONDARY SCHOOL SCIENCE	280
Problems in the measurement of science—Van Wagenen Reading Scales, General Science, Scales A and B—Downing Range of Information Test in Science—Grier Range of In-	

formation Test—Ruch-Popenoe General Science Test—Dvorak General Science Tests—Powers General Chemistry Test—Rich Chemistry Tests—Iowa Physics Test—Hughes Physics Scale—Columbia Research Bureau Physics Test—comparison of tests in physics—other tests in secondary school science—conclusions—materials needed—supplementary list of tests—selected references	
CHAPTER XV. FOREIGN LANGUAGES	299
Objectives in the study of foreign languages—American Council Alpha Tests in French, German, Spanish, and Italian—American Council Beta Tests in French, German, and Spanish—Columbia Research Bureau Tests in French, German, and Spanish—Ullman-Kirby Latin Comprehension Test—White Latin Test—other Latin tests—conclusions—materials needed—supplementary list of tests—selected references	
CHAPTER XVI. GENERAL ACHIEVEMENT TESTS	316
The place of the general achievement test—New Stanford Achievement Test—Sones-Harry High School Achievement Test—other general achievement tests—conclusions—materials needed—supplementary list of tests—selected references	
CHAPTER XVII. INTELLIGENCE TESTS	330
The nature of the intelligence test—types—some representative group tests—performance tests—aptitude tests—materials needed—supplementary list of tests—selected references	
CHAPTER XVIII. STATISTICAL AND GRAPHIC METHODS	346
Widespread use of statistics—statistical and graphic methods—measures of central tendency—measures of deviation—coefficient of variation—correlation—the column diagram and the learning curve—frequency curves—selected references	
INDEX	391

EDUCATIONAL MEASUREMENTS
AND THE
CLASS-ROOM TEACHER

EDUCATIONAL MEASUREMENTS AND THE CLASS-ROOM TEACHER

CHAPTER I

REASONS FOR EDUCATIONAL MEASUREMENTS

Traditional methods of grading.—Teachers, more than most persons, have been privileged to pass judgment on others with respect to a very significant characteristic: namely, the extent of one's knowledge and learning. Not only is this an important quality in a person's total make-up, but it is one concerning which many individuals manifest a marked sensitiveness. This sensitiveness is found especially in parents who, of course, desire the knowledge and conviction that their children are of a high level. It is not altogether strange, therefore, that these parents have been unwilling to admit the correctness of a teacher's judgment of their children—a judgment expressed in terms of "marks" or "grades"—when the rating of the children does not coincide with their own conviction; and so the teacher is accused of unfairness and inability.

The teacher has arrived at her estimates by means of recitations and questioning, both oral and written. The answers to these questions she has evaluated; and on the basis of these values she has determined that a

certain pupil merits an "A", or a "B", etc. Such measures and estimates coming from the teacher are highly subjective; that is, they are the result chiefly of individual construction, opinion, and sometimes guess-work, wherein the personal equation has had more or less free play. Under conditions such as these it is inevitable that teachers' estimates have at times been in error, and that personal bias should have operated, with or without the awareness of the teacher.

It is only natural, therefore, that marked dissatisfaction with unreliable methods of interrogation and evaluation should have arisen not only among parents but among some teachers and other educators as well. Thus during approximately the last twenty years we have witnessed the development of measures of school learning known as objective subject-matter tests, the development being particularly pronounced during the last dozen years. These tests, as the term *objective* indicates, are intended to eliminate the influence of personal and individual bias and opinion in the measurement and scoring of school achievement. Yet, in spite of their widespread and increasing use, their nature and function are not clearly understood by all teachers, supervisors, and superintendents; nor by parents. It is appropriate, therefore, in opening our discussion, to review briefly the reasons for the construction of better educational measures than we have had in the past.

The need for more adequate methods of grading.— Three questions arise which must be answered in order to make these reasons clear: First, what need has the teacher for any concrete standards for measuring the work of the class-room? Second, why are the methods of

the past not adequate? Third, has any real progress been made toward a solution of this problem which may afford relief to the class-room teacher? These three questions are closely related and will be taken up in order.

THE TEACHER'S NEED.—The teacher needs some sort of concrete standard of measurement in at least three inevitable relations: (a) the need in relation to the public, particularly as represented by the parents of the pupils; (b) the need in relation to the superintendent of schools, the supervisor, the principal, or other officer in immediate charge of the teachers; (c) the need for standards in the relation with the pupils themselves, involving the teacher's own guidance in the class-room.

Need of standards in relation to the public.—The teacher must have some sort of means for acquainting the parent with the attainment of his child. Fathers and mothers expect this report as a matter of course, and if it is not forthcoming, insist upon it. Thus one reason for the failure of attempts to do away with school marks and marking systems has been the refusal of parents to agree that measures of progress are unnecessary. Doubtless the typical parent has been improperly educated in the application of such standards, especially to his own children, but this makes all the more important a kind of mark which will carry with it a clear idea of the reasons for the child's success or failure. Very few parents are content with a system which designates simply *Passed* or *Failed* as the teacher's verdict on the work of the month or semester. Parents generally wish to know something of the relative positions of their children; whether the children are working somewhere

near their mental capacity; whether they are diligent; whether they try to succeed; and something of the general character of their attempts.

The parent is not the only person interested in the work of the teacher. The mother's clubs, and other women's organizations, parent-teacher associations, the board of education, are all bodies which are interested in the work of the individual teacher as well as of the school system in general. The teacher is often asked to appear before one or another of these bodies, sometimes to explain the work of the class for the information of the public, sometimes on the defensive, to meet criticism, sometimes as a member of the body to extend the influence of the school. In such relations, easily explained and readily understood standards are most essential to give point and emphasis to her report. If she can show by definite comparisons that her work in such a subject as arithmetic, reading, history, or any other, surpasses the average as set by the national standard for that subject, she scores a distinct triumph and wins approval for her school and her city. Particularly is such a result a happy one if her teaching has been criticized or attacked; for when she has a standard with which to compare her work specifically, and to demonstrate her success without cavil, she can silence critics in the most satisfactory way.

Need of standards in relations with supervisory officers.—The work of the supervisor is to improve the results of the class-room teacher. This holds likewise for the relations of the superintendent, the principal, and the special supervisor to the teacher. Wise supervision is directed to the strengthening of the teacher,

and to capitalizing her strong points, not to the criticism of the instructor's weaknesses, or to an attempt to bring out her failures in bold relief. Her successes are to become habitual, her failures to become negligible. But in the past there has been much difficulty in the way. Teachers have claimed successes which the supervisor could not recognize for lack of an easily applicable measure of success. Superintendents have set up standards for judgment which the teacher could not understand; and, indeed, the supervisor has frequently not deigned to explain to the teacher the standard by which the judgment was made. One reason for this has been that the standard existed only in the mind of the superintendent and had not been defined in concrete terms in his own thinking. He might even go so far as to say that his standard was indefinable—that good teaching and bad teaching were to be *sensed* but that the difference could not be expressed in definite terms. Under such supervision, the teacher was largely helpless; she knew whether her work was approved or not, but she was not conscious of any standard by which this result was reached. So the need has been felt very definitely for standards by which the ideas of the supervisor might be passed on to the teacher in terms of objectives to be striven for by the teacher with a clear knowledge of the goal to be attempted; and by which there could be a clear understanding by both parties of the final decision as to success or failure to reach this goal. Such standards are essential if there is to be the sympathy between supervisors and teachers necessary to the best results. This sympathy has not obtained in the past in a large number of our communities; and there has grown up an

antagonism between the supervisors and the teachers which has retarded the work of the schools. If concrete measurements were used, much of this antagonism would disappear. The supervisor would not only be welcome to the class-room but would be sought after and invited.

Need for standards in relation to pupils.—Even though the teacher did not have to justify her work to parents or to supervisors, she would nevertheless be faced with the necessity of satisfying herself as to the efficacy of her methods, and as to her success in handling individual pupils. When each pupil has a specific goal set for him to approach, he will work much more intelligently and frequently more willingly than otherwise. When he understands the reason for assignment of a recitation mark, or a semester evaluation, he will be less likely to charge his success or failure to the whim or partiality of the instructor. If this standard is sufficiently clear, he will be able to assign his own mark without the evaluation of the teacher being involved at all. He will also understand his own mark in relation to those received by his fellow pupils. Thus, the better the standard, the more likely is the teacher to have co-operation and zealous effort from the pupil. This consideration suggests an important psychological factor: namely, that if learning is to be most effective and achieved in the most economical manner, it is essential that the incentives or motivations be such as to effect a dynamic, participating attitude on the part of the pupil. That is to say, self-activity on the part of the pupil is essential. With a clearly conceived goal or standard in view, the desideratum is far more likely of attainment

than through vague and mysterious generalities of teacher or supervisor.

Again, the teacher should have standards which will indicate adequately the real differences between individual pupils, for the present emphasis upon the study of individual differences between children demands a definitely objective measure of these varying capacities. No other field of investigation in recent years has been so potent in developing and modifying the philosophies and practices of education as the psychology of individual differences, which owes its significance, in large measure, to the growth and improvement of tests of mental ability and tests of achievement in school subjects.

The matter of a final determination of the pupil ready for promotion, or of the child to be retarded or accelerated, demands standards based upon performance, not opinion, which will be of genuine assistance in deciding this most important question.

More than that, the instructor frequently wishes to know the answer to the question: "Have I got the results that I should have obtained? Is my class up to standard?" Many a teacher has worked a lifetime without actually knowing whether her work has been really good or really bad, superior or mediocre. Surely every teacher wants some definite information as to the true quality of her work. And she would much prefer to find this out for herself than to have it discovered for her by a supervisor or inspector. If she can have proper standards in her hands, she can do this. The need is very apparent.

A comparison of the results of tests in her various

classes with the norms made by classes generally over the country, or in communities similar to her own, as shown by the measurement of large numbers of children of all types, will give her a basis with which she can compare the result of the same or equally difficult tests given at an interval of a month or longer. Thus she will have a definite measure of the effectiveness of her teaching. This will enable her to measure progress of the pupils, and to know whether she has done as good a piece of teaching as she had supposed. She may find that she has really been succeeding, where she could not realize that progress was being made; or she may find that the marked improvement which she had noticed was not so great as she should have obtained. But in time she will be able to estimate her own proficiency more capably than was ever before possible.

INADEQUACY OF TRADITIONAL METHODS.—The methods of the past are not adequate to meet these various needs because they have not been based upon sufficiently sound scientific or pedagogic methods. Let us examine the methods most frequently used. They have been (a) examinations, (b) marks based upon recitations and tests, and (c) opinions of teachers and supervisors.

Inadequacy of examinations.—The examinations as ordinarily given in the past have been faulty in one or more of the following particulars:

(1) They have been constructed without a clear understanding of their purpose. In general, the purpose may be to test the pupil, or, at times, to test the teacher through her students. In many tests neither of these purposes has been clearly defined. If the object of the examination is to test the pupil, is it intended to dis-

close how well he has done specifically assigned work, or to find out how much he knows concerning a subject which he has been studying over a prolonged period? He might fail on the latter but pass the former. Is the examination designed to test memory, reasoning power, retention of information, or some other function? In other words, we may ask the student to reproduce factual data; or we may require that he interpret facts, or attempt the solution of a problem new to him; or, again, we may demand creative effort of him. It should be clear that any or all of these requirements might be blended in a single examination in a variety of ways. Even a cursory inspection of the process of examining reveals problems of great difficulty, and it is necessary, therefore, that the purpose of a test be clearly conceived and the results consistently interpreted. Furthermore, when examinations have been set by persons not in control of a specific class, they have frequently failed of their purpose because the teacher has not understood the intention in the mind of the examiner, and the pupils have consequently not been trained to meet the particular idea suggested. This has been unfair to teacher and pupil alike.

(2) They have not been constructed so as to make possible an accurate rating when corrected. This has reference to the several items in the examination. In fact, there has scarcely ever been any real agreement among teachers with respect to proper standards for rating the various parts of a paper. If there are ten questions what should be the allotted value of each? If this point is raised with reference to a specific paper in mathematics, or history, or English, or any other

subject, it will be found that even among a group of experienced teachers there will be a decided lack of uniformity in the evaluation of each question. This fact was given prominence by the investigations of Dearborn, Starch, and others¹ who were among the first to undertake to demonstrate the lack of reliability in teachers' judgments. A striking and well-known example² is found in the marking by 114 teachers of a single examination paper in geometry, each teacher scoring independently. Now, mathematics is regarded as a relatively concrete subject, in which the subject-matter is well defined; yet the marks of these teachers varied from 28 to 92. This is by no means an isolated instance. How much variation and difference of opinion may we expect, then, in the scoring of examinations in history, English, geography, and others where definition of materials and aims of instruction are not so clearly defined as in mathematics? When Question 1 is valued at ten points by one teacher, and at eight points by a second, and at three points by a third, all on a scale of 100, it will be seen that the pupil failing to answer this question only, will have a mark varying from 90 to 97. If such variations occur not on one question alone but on all questions in the examination, we find at least one

¹ Dearborn, W. F., "School and University Grades," *Bulletin of the University of Wisconsin*, No. 368. High School Series, No. 9. June, 1910. 59 pp.

Johnson, F. W., "A Study of High School Grades," *School Review*, 19: 13-24 (January, 1911).

Starch, Daniel, "Marks as Measures of School Work," and "A Sample Survey of the Marking System in a High School," in *Educational Measurements* (New York, The Macmillan Company, 1916), Chapters II and III.

² Starch, D., and Elliot, E. C., "Reliability of Grading Work in Mathematics," *School Review*, 21: 254-259 (April, 1913).

reason for the differences in scores for the paper in geometry, as reported above. Surely examinations are inadequate standards when there is no more accuracy in their framing than this!

(3) They are not corrected accurately even when made properly. Whenever papers have been sent to groups of teachers to be corrected according to given weights, but without further instructions, the variations have been so striking as to reveal the greatest inaccuracies in marking generally. This holds not only for such subjects as composition, which is supposed to be largely evaluated by judgment, but in mathematics, in which judgment in marking is not presumed to be an important factor. If Example 1 on an arithmetic or an algebra paper is weighted eight points for a perfect answer, shall one point, or all eight, be deducted for a single error in computation, the principle of the work being correct? Teachers are not by any means agreed on this point; and thus a paper with this single error may be marked anywhere from 93 to 99 points on a scale of 100. When errors occur in more than one question, this range is, of course, greatly increased. One does not have to go any farther to show the lack of agreement resulting from this failure to evaluate errors on a common basis.

A more astonishing fact has been brought out by experiment³ in that papers corrected by groups of teachers have been submitted to them for re-correction after a considerable lapse of time, and it has been demonstrated that the same teachers will not correct the

³ Starch, Daniel, *Educational Measurements*.

same papers twice alike. The variation of markings given under these circumstances is indeed surprising.⁴

Enough has been said to show that the examination in itself has been a most inadequate standard. Of all the old type examinations probably the most satisfactory is that which is set by the teacher herself to test her own class; but even this is not always suitable; for how many times we have heard teachers say: "I gave my class an examination to-day, and they did so badly. I thought they knew the work perfectly, but they made terrible mistakes," or, "I made my examination too easy, I think; everybody passed."

It is presumed that the experienced teacher, in setting examinations, has a definite plan and purpose in mind, by which the paper is constructed with direct relation to the teaching. But remarks such as those quoted indicate no such idea. In fact, until recently the matter of constructing examination papers has never been made the objective in educational classes, and, so, very few teachers recognize that there is really a scientific basis for making such papers.

Inadequacy of marks based on recitations and tests.—Many teachers have realized that examinations given at the close of work periods, as at the close of the semester, are not accurate measures of the pupil's ability, and have, therefore, either discarded the formal semester examination, or modified it by marks based upon the daily recitation and the quiz. Without doubt, such marks are more accurate measures than single examinations, but they are inadequate as really accurate standards for evaluation of the pupil's achieve-

⁴ See bibliography at end of chapter.

ment. There are several reasons for this that immediately suggest themselves:

(1) They are fragmentary. Unless the teacher is so great a slave to the record book that every pupil response is recorded, and this, of course, means inferior teaching, the record is after all only a partial one, and is open to dispute as to accuracy.

(2) They are not evenly balanced. The number of pupil responses is not the same for all; one great criticism of teachers by pupils is that some children are called upon in class much more frequently than others. Nor are the easy and difficult questions evenly distributed among all pupils, so that daily marks do not represent the same level of achievement for all pupils.

(3) Where the written quiz is used, the same objections arise as for the more formal examination.

(4) Just as there is a great variation in marking papers, so there is a great variation in evaluating oral responses.

(5) The art of questioning, and the purposes of the question, are not much better understood by many teachers than is the science of making examination papers. This again makes the recitation an unsafe guide for marking.

To be sure, the average of a number of marks may conceivably be more accurate than any single mark either of recitation or examination. But the teacher cannot safely rely upon an average to correct single inaccuracies, for the average of unreliable scores is more reliable than the individual score only when the errors are symmetrically distributed on both sides of the true measure. This is very unlikely to happen in assigning

class marks. So the law of averages can not be invoked to correct marking errors.

Inadequacy of opinions of teachers and supervisors.—Results of examinations and averages of class-room marks as bases for evaluating work of teacher, class, or individual pupil have for a long time been recognized by keen observers as inadequate measures. But the conclusion reached during a period of many years was that no more adequate measures could be made, and that the safest guide in applying these measures was the experience and good judgment of the teacher and superintendent. Accordingly, in many cases where the faulty examination produced the failing pupil, the teacher substituted her judgment as to the pupil's ability and passed him to the next school grade, regardless of his seeming failure. Even where the class record was not good, the superintendent declared that, since the teacher had used methods of which he thoroughly approved from his own experience, he would promote the teacher despite her apparent failure. Doubtless, a teacher of long experience who had developed unusual skill in diagnosis might employ such methods with comparative success. The judgment of a good superintendent in evaluating work of teachers and classes is not to be thrown aside as of no value. But while a few persons unquestionably have such qualities, the rank and file of teachers and supervisors are lacking in experience. Without experience on which to base judgment, keenness in diagnosis is not likely to command much respect, where it is based solely upon one's opinion. Such judgments are generally spoken of as subjective; that is, they are dependent upon a standard

originating with the person giving the opinion, and upon his own individual conclusion, reached through the consideration of data developed from his own experience or his own personal tests. A subjective judgment can never have the force of an objective conclusion, for the objective standard is based not upon one's own experience, but rather upon data obtained by methods which have developed from the combined experiences and judgments of many. The objective conclusion, therefore, is more concrete in form and substance than the subjective estimate. The tendency of the subjective conclusion is to disregard objective data; the objective conclusion is directly dependent upon objective data. The subjective judgment is, therefore, always open to criticism, doubt, and successful attack. The position of the objective conclusion is reversed; it is based upon data scientifically assembled and standardized; its position is, therefore, firm.

PROGRESS IN EDUCATIONAL MEASUREMENT.—As a result of the experiments and studies made during the past fifteen years, the third great question may be answered categorically. Great progress has been made, and is being made, to solve the problem of proper measures of educational attainment. Instead of dying out or losing force, the movement is becoming so generally accepted in principle, that it has passed the stage of discussion, and for this very reason is not occupying the space in controversial literature that it did fifteen or twenty years ago. At that time, the question was: "Is it possible to measure any educational achievement objectively?" Now the only question is: "Is there any educational achievement which can not be measured

objectively?" The alert teacher now finds it possible to measure the greater part of elementary school attainment and a large portion of the secondary school curriculum by well standardized tests or scales. Every year more and better measures are being produced; those already in use are being perfected; and the field of application is being still farther extended.

In Chapter II we shall discuss the nature of these measures in the light of the inadequacies of traditional methods and classify the types of measures now in use.

SELECTED REFERENCES

Buckner, C. A., *Educational Diagnosis of Individual Pupils* (Teachers College, Columbia University, 1919).

Carter, R. E., "Correlation of Elementary Schools and High Schools" (*Elementary School Teacher*, Vol. 12, p. 109).

Comin, R., "Teachers' Estimates of the Ability of Pupils" (*School and Society*, Vol. 3, p. 67).

Dearborn, W. F., "School and University Grades" (*Bulletin of the University of Wisconsin*, No. 368. High School Series, No. 9. June, 1910).

Inglis, Alexander, "Variability of Judgments in Equalizing Values in Grading" (*Educational Administration and Supervision*, Vol. 2, p. 25).

Johnson, F. W., "A Study of High School Grades" (*School Review*, Vol. 19, p. 13).

Kelly, F. J., *Teachers' Marks* (Teachers College Contributions to Education, No. 66).

Rugg, H. O., *The Teachers' Use of Statistical Distributions in Giving School Marks, A Primer of Graphics and Statistics for Teachers* (Boston, Houghton Mifflin Company, 1925. Chapter VI).

Starch, Daniel, *Educational Measurements* (New York, The Macmillan Company, Chapter 2).

Starch and Elliott, "Reliability of Grading High School Work in English" (*School Review*, Vol. 20, p. 442). "Reliability of Grading Work in Mathematics" (*School Review*, Vol. 21, p. 254). "Reliability of Grading Work in History" (*School Review*, Vol. 21, p. 676).

CHAPTER II

WHAT CONSTITUTES A STANDARD MEASURE

Factors involved in standard tests.—The discussion of Chapter I has set before us very definitely the need for measures of class-room work which will be free from any suspicion of unfairness in construction or application, which may be relied upon to tell the teacher more nearly the truth about the progress of the pupils, which will indicate her own success or failure in presenting the work, and which are quite free from the whim or prejudice of either teacher or superintendent. To meet these needs a number of different sorts of measures have been devised. As yet, although remarkable progress has been made, the need has not been entirely met; but enough has been done to show that the making of standard measures is possible, and that although absolute perfection has not resulted, the measures now in use are so far in advance of the imperfect methods of tradition that no teacher is justified in remaining in ignorance of their scope or of their application.

This chapter will be devoted to a discussion of the factors involved in making such tests of achievement, and of the various classes of tests and measures available. This discussion will involve three phases: (a) Dif-

ferentiations from traditional methods; (b) necessary specifications; (c) resulting types of measures.

Differentiations from traditional measures.—To be entirely satisfactory, our new measures must be free from the objections suggested in Chapter I as holding for the methods of the nineteenth and previous centuries. They must avoid the vagueness and indefiniteness of the old type of test; they must be based upon a definite knowledge of the element to be tested, and must be framed to test that element particularly. Each part of the test must have a definite value, ascertained by scientific methods. No part of the measure can be left to determination by opinion of the examiner, either in its origin or its application. So far as humanly possible, the subjective element must be eliminated.

Now these various requirements are at present possible of fulfilment in the main, whereas in the nineteenth century the necessary means were largely lacking. Only in the last quarter century have educators applied to educational products the mathematical methods used for many years by biologists, astronomers, and other scientists. Thus, the last twenty-five years have created a scientific side of education which is an entirely new development and which makes possible many things before thought beyond our reach.

Educational experiments are not now considered valuable unless they are carried out under scientific principles, and controlled and interpreted by scientific method. Our measures are therefore to be based upon scientific experiment and mathematical interpretation. If the element of judgment is involved, it must also be subject to scientific scrutiny, and unless it can meet such search-

ing inquiry, must be discarded as worthless. So even where a subjective factor may be involved, it becomes, in the light of scientific method, objective in its application. We then feel justified in speaking of all measures, tests, and scales, evolved under scientific principles, as objective measures both in evolution and in application. No longer can the opinion of a single individual, or limited group of individuals, have weight in framing measures of scientific value.

In appearance, some of the new tests resemble exactly the old subjective examination, and the uninitiated teacher can not discover the value of the new over the old. The value lies in the fact that the new in all of its parts has been subjected to scientific scrutiny which has established the fact that it is free from the faults of the old; it is known definitely what it is intended to test; every part has a specifically determined value; it is not the result of individual opinion or whim; in correction of answers to the questions, little or nothing is left to opinion; every answer has been weighed beforehand, and provision made for evaluation of variations from the precise response desired. The personal equation no longer holds in framing or correcting this new paper, for it is truly objective.

Necessary specifications of the objective measure.— In order to avoid the weaknesses of traditional measures, and to meet modern scientific requirements, measures of class-room achievement must now conform with certain definite specifications. The most important of these are as follows:

NEED OF DEFINITE AIM.—The aim of instruction in the subject tested must be clearly envisaged. In teaching

addition, the aim may be to render the subject of instruction competent to react automatically to a certain situation; in other words, to drill him until he is letter perfect in various number combinations. The measure which tests these automatic reactions is then the essential measure, and must be made with this aim in view. But if the aim is to teach reasoning in arithmetic, through the medium of combinations involving addition only, the new aim must be held in mind in testing the pupil or class. The measure will then vary definitely from the first. A measure of addition cannot be made to conform to our new standards, unless it is known what aim is involved in the teaching. This principle holds for all subjects in the curriculum. Both the investigator in collecting materials for the making of the measure and the class-room teacher in using the measure are at fault unless they have this first essential, the aim of instruction in the subject, definitely in mind.

It should be clear that it is not the function of an objective test to *define* the aims and purposes of a course of instruction in any subject. The question of values and the justification for the inclusion or exclusion of subject-matter are not and can not be determined primarily by an examination. The test may contribute toward a final evaluation of materials; but its principal task is to measure accurately and completely those aims and objectives already determined.

MATERIAL MUST BE REPRESENTATIVE.—After the aim is clearly in mind, there must next be a selection of representative material which will develop and expand this aim. The materials entering into any subject must be comprehensive enough to make the realization of the

aim possible, without involving extraneous elements which confuse and complicate the result. This is a principle of pedagogy inherent in teaching the subject, and the same principle must be kept in mind in making the measure of the subject. Right here we may again emphasize the fact that objective measures are to be used to evaluate the results of teaching, and that for the attainment of this end it is essential that the measures be constructed with a clear understanding of the best pedagogical practices. A scale or test which involves materials rejected by competent teachers on the ground of poor adaptation to instruction is indeed a poor scale. The test which is so limited in the scope of material covered that it does not give opportunity to demonstrate the result of broad-gauge teaching, is not adapted to show the possibilities of excellent class-room instruction. The material must be in every way representative. It must be comprehensive, properly selected in scope and difficulty, adapted to the best teaching method, and suited to the best realization of the aim of the subject.

QUANTITATIVE ELEMENTS AND METHODS OF MEASURING THEM.—That subject-matter is best suited to measurement by objective scales or tests which contains the greatest quantity of objective elements. The earliest criticisms of objective measures were based upon the statement that all teaching is so largely qualitative that no exact measures were possible. The critics said: "How can one measure appreciation of a beautiful poem?" They felt that such matter of instruction held but few quantitative elements, and so was unsuited to measurement.

Obviously, it is easier to construct a measure of

quantity than one of quality. Naturally, then, we feel more nearly certain of the results obtained with a measure dealing with quantities; and we therefore accept with but little hesitation the measures of proficiency in spelling, of the fundamental processes in arithmetic, of descriptive data in geography or history, and of the rate of reading. But measures of proficiency in penmanship, English composition, interpretation of history, are not so readily received as valid. However, as the study of measurement has gone on, the surprising thing, to many, is that the quality of intangible matter has been shown to lend itself to rather accurate quantification, by which degrees of excellence are disclosed.

By skilful questioning, the appreciation of a poem can be measured to a point where the teacher of literature feels quite justified in recording a subjective mark to indicate the relative interest, enthusiasm, esthetic understanding, and other facts of appreciation shown by different members of the class. A good measure, therefore, must be based upon a validated quantification of these factors; and the author of the measure succeeds or fails according as he is able to separate the quantitative from the non-quantitative elements of the subject.

When there is a quantity of any thing, it is possible to measure that quantity. Our familiarity with ordinary measuring scales is the very reason for our failure to understand the problem that confronted primitive man in devising proper measures. It is easy to imagine early men saying: "There is certainly such a thing as heat. But it is subjective, a thing of the senses, and varies in subjective sensation with each individual, according

to his susceptibility. No measure can be found to scale this intangible sensation!" In the light of such reactions, how marvelous does the thermometer become! So it is with the products of the class-room. Our methods of measurement are based upon the same principles that were used by scientists in devising methods for measuring intensity of light, electric current, flow of gases, and such intangible things.

The arbitrary zero point on the thermometer is perhaps suggestive of a level on the school product scale which shall differentiate between passing and failing. But educational measurements in general are made on scientific principles which rule out such arbitrary assumptions; rather are the divisions based upon actual achievements of large numbers of individuals. Our methods are, therefore, the outgrowth of experiment with and trial of proposed measures to determine those factors which can be scaled by means of relatively definite procedures, with elimination of elements which do not lend themselves to such specific treatment.

A varied sort of measure results; we may measure reaction to certain situations on the part of the pupil, as when he is asked to perform given tasks which are then evaluated according to fixed standards; we may compare specimens of his work done under normal conditions with models containing the same or similar elements; we may evaluate oral responses; we may use his written responses; we may measure sensori-motor as well as psychical factors, especially in connection with manual dexterity involved in such manipulatory subjects as penmanship, typing, telegraphy, and the like. In these last, a diagnosis of manipulation is necessary

in order to make a satisfactory measure. In general the builder of a test endeavors to analyze the elements of mental activity involved in the various situations investigated. It is this factor which sometimes complicates the result of the work to an extent that makes care necessary in evaluating results, and frequently makes a reëxamination of the individual important before passing judgment or drawing hasty conclusions. The warning is given here, as it will be strongly emphasized hereafter, that the study of the individual child must always be made in the light of repeated and varied measures, rather than that his status or future be determined on the basis of a single measure or single set of measures.

With reference to the whole method of measuring quantitative elements, we must stress the fact that, so far as possible, measures must be based upon the possibility of isolating these elements from concomitant factors, and that the validity of measures is affected principally by the extent to which this is accomplished. That is to say, a test, to possess the highest validity, must measure what it purports to measure, without involving extraneous factors or processes. Obviously, unless this condition is satisfied, the results of a test are ambiguous.

STANDARDIZATION OF RESULTS.—The chief difference between certain forms of examination, such as those made up by examining boards, and objective measures of the sort described herein, is that the latter are standardized, while the former are unstandardized and likely to be unstable in nature. The examination constructed by a large committee, each member of which actually participates in its composition, is not subjective in the

sense that it is the product of some one person's opinion; but, as the result of collaboration of presumably competent judges, it is felt to be free from individual bias or prejudice. Yet, at the best, it is untried in the sense that it represents only a consensus of opinion, and not a result of a scientific experiment or computation.

Standardization of a test is arrived at by actually administering it to a large number of individuals within specified age or grade limits, and by subjecting the results to careful statistical scrutiny. The process is not a simple one, frequently involving extensive adjustment and alteration before the final form is satisfactorily established.¹ Thus, a test which has been standardized has been proved of a certain validity, a fact which makes it of tremendously more value than the test which is dependent for its significance solely upon the opinions of the persons who have devised it.

The standard of actual achievement made by groups of pupils taking a given test is called the *norm* of achievement for that test and that group. If a norm is desired for children of a given grade generally, this norm would, theoretically, be the result obtained by giving the test to all children of that grade in the United States and then taking the norm to be the score of the average,² or, presumably, the normal child of the group. Obviously, it is impossible to give a test to *all* children of a specified group; instead, representative groups,

¹ For a discussion of the construction of mental and educational tests, see Part IV of *Tests and Measurements in High School Instruction*, by Ruch and Stoddard, Yonkers-on-Hudson (World Book Company). Also *Statistics in Psychology and Education*, by Garrett, especially pp. 101-115. (New York, Longmans, Green & Co.)

² That is, the mean or median. A discussion of averages, or central tendencies, will be found in the last chapter.

called random samplings, must be the source of our information and data. Fortunately, mathematical statistics have demonstrated that if data are obtained from a large enough number of representative children in different parts of the country and in a variety of communities, the results are valid for all children of the specified group. It should be carefully noted that we have spoken of *representative* children as constituting the random sampling. It is not enough merely to have a large number; but this number must be properly distributed; otherwise we may get what is known as a selected group. In such an event the norms are not generally applicable.

When norms are established, the interpretation of scores with respect to them may be of one or more kinds. We may simply refer a given score to the age or grade norm, as the case may be; or we may refer the individual's position to the other members in the group in terms of either percentile rank or deviation position. Both of these latter statistical devices are designed to give an individual's status with precision and consistency.³

The norm, as used in education and psychology, refers to the results of *actual* achievement. However, the *objective* set before the teacher need not and does not necessarily coincide with a given norm, for educators may reasonably expect that the standard of achievement should improve under improved methods and facilities. Or, on the other hand, it is conceivable that the norm may be too high for a certain community laboring under

³ For a detailed account of the calculation of these indices see any of the recent standard texts in statistics, such as Holzinger's *Statistical Methods for Students in Education* (Boston, Ginn and Company), and Garrett, *op. cit.*

serious handicap, such as consistently poor pupil material, a too short school year, undifferentiated classes. The *standard* is, therefore, likely to be higher, and at times lower, than the *norm*, and is to an extent theoretical in assumption and application. This distinction between norm and standard will be kept throughout the book.

The teacher must keep in mind that the norm is the measure which she is to use most commonly in connection with standardized tests. But norms for the entire country or for any large portion of it should not be over-evaluated in any particular instance. At times, it is much more significant for a teacher or administrative official to know the central tendency of a certain grade only in communities comparable to his own in point of size and pupil population. It might even happen that the most significant datum will be the central tendency for the grade *within* a single school system, by which the effectiveness of teachers, methods, and curricula can be gauged. In other words, norms for the nation, state, and local community are all valuable; but which index will be meaningful in a specific situation will depend upon the problem involved and the purpose to be served.

The criteria of a good test.—By way of summary we may set down briefly the criteria which have been regarded as most important in the selection of a standardized test. Broadly, these criteria may be classified under two general heads: namely, statistical and curricular.

STATISTICAL CRITERIA.—Under the first, statistical, the following considerations may be enumerated:

(1) *The test should be objective.* We have already discussed the meaning of objectivity, so that further details are not necessary here. Those responsible for the selection of a test will, of course, investigate such matters as objectivity of scoring keys; adequacy of directions for administering, scoring, and computing results; equivalent forms to eliminate practice effect as far as possible.

(2) *The test must be reliable;* that is, it must give the same or very nearly the same results on re-tests. It is clear that if a measure ranks a group now in one order and then in another, it is hardly a useful instrument. The reliability of his results will generally be stated by the author of the test.

(3) *The test should measure over a rather wide range of ages and grades.* Unless this condition is satisfied, the value of a test is decidedly limited, for wide comparisons and accurate predictions are otherwise not possible. When this condition is satisfied, the degree of overlapping in ability from age to age, or grade to grade may readily be seen. The importance of such information for the purpose of dealing with individual cases, or for a better understanding of some aspect of a school problem is obvious.

(4) *The test should provide reliable norms.* If the norms are inadequate or unreliable, the measure loses a considerable portion of its value. As pointed out earlier in this chapter, the nature of the norms and their significance for any school system or teacher will vary, depending upon the problem and circumstances in each situation. But whether norms be desired for the nation, the state, or a restricted community, whether

for ages or grades, the test should be selected with one's peculiar problem in mind. Merely administering tests and comparing results with any sort of norm or standard serves no justifiable purpose.

CURRICULAR CRITERIA.—Under the second division, curricular, we may list the following criteria:

(1) *The test should be valid*; that is, it should measure the ability or subject-achievement which it purports to measure. For example, if a test is designed to measure reading comprehension, it should not involve to an appreciable extent the ability to write well or rapidly. If it does, then the score is much less meaningful, for it does not show how far the result has been influenced by writing ability. Nor, for instance, should a measure of arithmetical ability make such demands in reading that the test results are a function of reading ability as well as of arithmetical ability.

(2) *The test should be consistent with good teaching practice*, and its sampling of important aspects of the subject should be adequate.

(3) *The test should be interesting to the pupils and not too long*. If interesting, the measure will arouse the active participation of the pupils; without this interest the results may be unreliable. If too long, the results suffer in reliability because of the pupils' fatigue and loss of interest. Further, too great length disrupts the teacher's schedule and requires time which might be spent in profitable instruction.

(4) *The test should have diagnostic value*. This is a necessary condition, of course, where the standardized measure is used primarily for a better understanding of a pupil's abilities or disabilities. It is more important

to know the special difficulties which contribute to failure than merely to know that a pupil is failing. A test may also have diagnostic value for a class as a whole, as, for example, in pointing out peculiar difficulties in a certain type of arithmetical problem; or, in detecting what letter combinations are not being mastered in spelling; or, in reading, to determine whether comprehension is being sacrificed for rapidity.

It is doubtful if any tests meet all these ideal conditions; but some approach them more closely than others. However, as educators define the aims of instruction more accurately, as the work in the field of standardized measures progresses, we may very reasonably expect a closer approach to the ideal.

Different kinds of measures.—Most that has been written so far has referred more particularly to what are known as tests or scales to be used in connection with the class-room to determine the degree of attainment pupils have reached in their regular daily work in such subjects as arithmetic, language, spelling, Latin, history, etc. Such measures are known as achievement or attainment tests. They are to be distinguished from tests of mental ability, or intelligence, for the latter are not intended to measure school progress, but rather the individual's capacity to perform what we are pleased to call the intelligent act.

SCALED MEASURES.—Measures of achievement are of two kinds: namely, scales and "tests"; or, scaled and unscaled. The former is a measure in which the items progress according to some definite quality, such as difficulty or esthetic value. In the measure scaled on the basis of difficulty, for example, each item is more diffi-

cult than the one preceding it. Furthermore, in the ideal situation, the increase in difficulty from any one item to the next should be equal to the increase from any other item to *its* next. That is, the increase in difficulty from number 5 to number 6 should be equal to the increase from 10 to 11. It will be seen that in a scale of this sort an individual's score generally represents the *level* which he is able to achieve in the given subject; for, in addition to being scaled, the good measure will be such that no pupil in the group for which it is intended will get a score of zero or a perfect score. A zero score should represent "not any" of the quality being measured; a perfect score should represent perfection in the quality. But two pupils might have zero scores yet differ in ability. So, too, several pupils achieving perfect scores might differ in the extent to which they are able to surpass the perfect score of the particular measure employed.

UNSCALED MEASURES.—The second type, the "test" or the unscaled measure, differs from the first in that the items are of equal or very nearly equal difficulty. In an instrument of this sort, the factor making for differences in individual scores, is the rate of performance, or what is commonly known as *speed*. Unlike the scale, these "tests" do not indicate the *level* of difficulty in the subject beyond which the pupil is unable to progress. The "test," however, is not to be dismissed because of this apparent inferiority. A number of items, different in kind, may be of the same general difficulty, yet in individual cases one *type* of item will be of greater difficulty than another. For instance, consider a "test" in multiplication. The items are so selected that all the

important combinations are included so that weakness in any given combination will be revealed. Thus it may be found that one pupil is experiencing unexpected difficulty where the figure 7 is involved, while another finds 9 a stumbling block. For the most part, however, teachers and administrators will find the scaled measure of greater value than the "test", all other considerations being equal.

Mental tests.—Tests of mental ability, however, are quite different in conception from achievement tests. The former are of two sorts, laboratory tests, chiefly of the sensori-motor and memory types, used in the laboratory of psychology to test individual reactions to a variety of stimuli, together with measures of certain anthropometric and physical traits. Among these may be numbered tests of strength or grip; quickness, accuracy, and precision of movement; visual and auditory acuity; range of visual apprehension, and the like. Of the more complex sort, there are tests of controlled and uncontrolled association; of rote and logical memory; of suggestion.⁴

A second type of test of mental ability is what is commonly known as the intelligence test, either group or individual. These, consisting of series of questions and problems, are administered in groups or singly, as the case may be, in order to determine the pupil's ability to perform tasks of a general character not connected, at least directly, with school work. In a later chapter we

⁴ The mental ability tests of the laboratory of psychology are well represented in Whipple's *Manual of Mental and Physical Tests* (revised edition) (Baltimore, Warwick and York). To this manual the reader is referred for a study of the various laboratory exercises and their significance.

shall have more to say with respect to the intelligence test and its meaning; but for the present we wish only to introduce the distinction of types.

During the past fifteen years the individual and group tests of intelligence have achieved rather remarkable prominence and widespread use, inasmuch as it has been possible to administer them in schools, rather than being restricted to the laboratory and the necessarily small number of individuals who can be studied under laboratory conditions. The intelligence test, as it is known to-day, has grown out of the work of the French psychologists, Binet and Henri, and later Binet and Simon.⁵ Their scales were brought to this country, revised and adapted for use with American children. Of the several revisions, perhaps the best known⁶ and most widely used is the Stanford Revision of the Binet-Simon Tests, published in 1916.⁷

As a result of the mental testing by group methods in the army during the World War, the group tests have grown in number and improved in quality. This, quite naturally, has resulted in their more widespread use than formerly, because of the large numbers who can be handled at one time, and also because of the relative ease of administering and scoring. Among the better known of the group tests are the Army Alpha and Beta, the Dearborn, the Haggerty, the National, the Otis, the Detroit First Grade Intelligence Test, the Terman, and the Thorndike. Although some of these are intended for nearly the same range of ages, others are designed

⁵ See Chapter XVII.

⁶ Other revisions: Herring Revision of the Binet-Simon Tests; Yerkes-Bridges Point Scale; Kuhlmann Revision of the Binet-Simon Scale.

⁷ See L. M. Terman: *The Measurement of Intelligence*.

for a limited group, as, for example, the Detroit for grade one and the Thorndike for high school graduates. There are, of course, other group tests perhaps equally as good as these mentioned.

Since the increased efforts in the measurement of intelligence, there has been considerable controversy with regard to the nature and definition of intelligence, and also concerning the question of what is really being measured by the intelligence tests.⁸ In a discussion such as this, where we are not concerned primarily with intelligence tests and their theoretical implications, we hesitate to enter upon controversial ground. Instead we shall at this point confine ourselves to matters on which there is rather general agreement. Whatever else the intelligence tests may do or signify, there are very good grounds for maintaining that they predict with a high degree of reliability the ability of children to do school work in general. It may be said with confidence that a child with a high intelligence quotient⁹—abbreviated IQ—(*high* meaning that he is well above the average in brightness) has the ability to do better school work than the average pupil. Likewise, it may be said that the child with a low IQ will be unable to do as good work as the average, other things being equal. But a high IQ does not *necessarily* indicate that the pupil will attain ranks relatively as high in arithmetic, reading, spelling, or

⁸ As an example, see: "Intelligence and its Measurement." A Symposium. *Journal of Educational Psychology*, Vol. 12, pp. 123-147, 195-216, 271-275.

⁹ We may, for the present, define the intelligence quotient as an index of a child's degree of brightness or of his rate of mental development. The normal IQ, theoretically, is 1.00 or, as often written, 100. An index above or below 100 indicates in varying degrees accelerated or retarded mental development, respectively. Great caution is necessary in the interpretation of the IQ.

any other specified subject. Although *in general* it may be said that intelligence tests and school subjects show a high correlation, and that the same is true of correlations among the subjects themselves, yet there enter other factors of sufficient importance to produce situations where high general ability may not be reflected in actual accomplishment. Not all of an individual's abilities are absolutely uniform; in a relatively small, but ever-present number of cases there may be exceptional abilities or disabilities; emotional factors may play a part; better than usual or worse than usual teaching might influence results; regularity of attendance, good or poor health, etc.—all these are forces to reckon with. Nor do the intelligence tests measure those other qualities which make for success, such as studiousness, industry, motivation, and the like. On the other hand, the low IQ may be accompanied by such a high degree of application and perseverance that the apparently handicapped child will in the long run surpass the more favored one who is not so well endowed with these desirable and necessary traits.

In concluding this general review of test types, we wish to emphasize the view that the term "psychological test" is too inclusive to carry a useful and specific meaning in most situations, unless the type is indicated. The test may be one of achievement, intelligence, personality, character; it may be of the simple sensorimotor type, or the more complex laboratory test of association, suggestion, motor skill, etc. If vagueness and misunderstanding are to be avoided, particularly in school problems and school experiments, tests should

be selected, used, and interpreted according to the purposes for which they have been designed.

SELECTED REFERENCES

Burt, Cyril, *Mental and Scholastic Tests* (London, P. S. King & Son, 1922).

Franzen, R. and Knight, F. B., "Criteria to Employ in Choice of Tests" (*Journal of Educational Psychology*, Vol. 12, Nov. 1921, pp. 408-412).

Freeman, F. S., "Influence of Educational Attainment Upon Tests of Intelligence" (*Journal of Educational Psychology*, Vol. 19, No. 4, April, 1928, pp. 230-242).

Garrett, H. E., *Statistics in Psychology and Education* (New York, Longmans, Green & Company, 1926).

Gordon, Hugh, *Mental and Scholastic Tests Among Retarded Children* (Board of Education, London, 1924).

Hollingworth, L., *Special Talents and Defects* (New York, The Macmillan Company, 1923).

Holzinger, K. J., *Statistical Methods for Students in Education* (Boston, Ginn and Company, 1928).

Roback, A. A., "Subjective Tests versus Objective Tests" (*Journal of Educational Psychology*, Vol. 12), Nov. 1921, pp. 439-444.

Ruch and Stoddard, *Tests and Measurements in High School Instruction* (Part IV) (Yonkers-on-Hudson, World Book Company, 1927).

Symposium: "Intelligence and its Measurement" (*Journal of Educational Psychology*, Vol. 12 pp. 123-147, 195-216, 271-275).

Terman, L. M., *The Measurement of Intelligence* (Boston, Houghton Mifflin Company, 1916).

Thomson, G. H., *Instinct, Intelligence and Character* (New York, Longmans, Green & Co., 1925).

Whipple, G. M., *Manual of Mental and Physical Tests* (revised edition) (Baltimore, Warwick and York, 1915).

CHAPTER III

THE PRACTICAL USE OF EDUCATIONAL MEASURES IN THE CLASS-ROOM

Introduction.—So much has been written and said about the value of tests as supervisory instruments, as aids to school surveys, and as research tools, that frequently class-room teachers have not thought of them as of especial value to the everyday teacher in solving her own problems of the class-room. Yet this is where the most vital and important uses should be found. This chapter will, therefore, be given over to a consideration of some of these uses. They seem to fall into two broad groups dependent on the two general types of tests, namely, achievement and intelligence.

Uses of Achievement Tests

For purposes of comparison.—Teachers frequently desire to know, and should know just how their classes will compare in attainment with other classes in the same school system, or in the state, or in the United States. Before achievement tests were worked out, this knowledge was practically impossible. But now administering a standard test in penmanship or Latin or arithmetic makes possible a comparison with the norms which have been determined for the country in general for that particular subject, or part of the subject, so

that one can easily see whether the class is doing what should be expected of it at its particular point of advancement, if it is a typical class.¹ In like manner, if norms have been determined for the state in which the work is done, or for the city or immediate community, a comparison which may be still more valuable can be made.

Comparisons with other classes within the system in which the teacher is working, within the same building, and even between different sections in charge of the same teacher, may also be made on a basis of the objective norms. Each of these comparisons has its own peculiar value in assisting the teacher to determine the relative attainment of any particular class at a given time. Where norms are not available except for the country at large, assistance in comparing special types of pupils may be frequently obtained by referring to various city, state, and county surveys, in which norms for large groups of children have been obtained. Also, by writing to bureaus of educational research in the larger cities in various parts of the country, norms may be had which will be of value for comparative purposes.

Another sort of comparison which is useful is that between the attainment of a class at the beginning and at the end of a semester's work, or at lesser intervals in the course of the semester. Such comparisons are decidedly stimulating to class work, if the interest of the class is aroused by having the purpose of the tests explained, and the goal of expected improvement set definitely before it. In such instances, the test itself is, of

¹ The importance of knowing the intelligence ratings, in addition to achievement norms for the purpose of evaluating the norms, will be considered in a later section of this chapter.

course, not set as the matter for drill, but is carefully kept from the pupils, to prevent any sort of unfair use of its content. This means especially that a pupil is not to be allowed to retain a copy of the test at any time that it is given.

Alternative forms of nearly all tests are available for comparisons of the sort indicated, so that if there is any suspicion that pupils are "coaching" each other, or have kept copies of tests, the alternative forms may be given. They are constructed so that they will be of equal difficulty with the originals, and so will be accurate for comparative purposes.

For diagnosis of teaching.—The comparisons just suggested are valuable in themselves by way of indicating whether classes are up to standard in their attainment. But it is even more important to be able to determine as a result of these comparative studies whether the teaching itself may not be improved. The result of the tests will show the effect of drill, for one thing. If the drill has been ineffectual, this will be evident. If too much time has been given to the mechanical aspects of a subject, the undue proficiency of the class will show that there has been a waste of time and effort along this line. This is what may be called *over-learning*. Some aspects of a subject may be learned within a reasonably short period. To review and to repeat constantly, therefore, is but to re-emphasize that which needs no more emphasis. This criticism applies, of course, especially to the more mechanical aspects of a subject. Where, however, there is constant *use* of learned materials, involving reasoning, judgment, and new applications, the

situation is altogether different; over-learning is then only a by-product, not the goal.

When the class shows marked strength along one direction and weakness along another, the instruction should, of course, be so modified as to remedy the weakness, while less effort will be spent along the lines where proficiency is evident.

An important caution, however, is necessary. It would be unfortunate if the use of standardized tests led to an over-emphasis of the mechanical aspects of any study. For example, in reading we can measure the number of words a pupil is able to read per minute; we can also measure his comprehension by finding how many questions he is able to answer about the selection read. But the tests do not measure the *effect* which reading has produced on the pupil; whether a given type of material makes him a more or a less desirable person. We can measure the scientific facts which a pupil has learned; but we do not necessarily know how far his contact with science has produced a scientific attitude. So, too, we can measure spelling ability; but ability to spell tells us little regarding the individual's ability to use the word correctly. Let it be noted, however, that we are not discrediting or rejecting the mechanical aspects of a subject; for example, too slow readers are handicapped; a fund of scientific information is essential for the development of a scientific attitude. But we do object to such a marked emphasis on the mechanical phases of a subject that teachers aim at little else, while other very significant features of instruction inevitably suffer.

Some tests are designed to examine the mechanical side of a subject, such as the Courtis tests in arithmetic which measure the fundamental processes of arithmetic. Others, however, are intended to go beyond mechanics, such as the Stone Reasoning Test in arithmetic. Through a wise use of both types of tests, it is possible for the teacher or the administrator to disclose wherein weakness may lie in a subject and to adopt a suitable remedy.

For evaluating methods of teaching.—Methods of presentation may also be evaluated by means of tests. Teachers and supervisors are seeking the best methods of instruction; and the most effective and accurate manner of evaluating methods is by the use of standardized achievement tests. The unreliability of tests constructed by individual teachers has been sufficiently discussed to make it clear that they are unsatisfactory in the solution of a question as important as this. Yet, before the advent of standardized tests, the inclusion of one method or the exclusion of another was dependent upon these highly subjective teachers' examinations; or, worse still, upon the bias and personal preference of teacher or supervisor. It is possible now, through properly conceived and controlled experimental studies to discover what methodological procedures are best for a given school subject or for a given group of pupils. Thus, it may be determined that spelling is best taught when the words are in a sentence rather than when isolated; or it may be shown that silent reading habits are superior to habits induced by oral reading, etc.

It will be clear that the teacher may discover much about her classes for herself; and that in coöperation

with supervisor and superintendent much valuable information may be had for the purpose of improving instruction.

For diagnosis of classes.—At the beginning of a school term the teacher wishes to know, and should know, a good deal about the proficiency of her classes in subject-matter and about the adequacy of their general preparation. She must learn where the weak points lie, in order to direct her work where it will do the most good; she desires to know the nature of the preparation her pupils have had for the work in general. For the purpose of the teacher's own orientation, standard tests are very valuable.

Such a test as the Barr Diagnostic Test for American History combines a number of factors into one test in such a manner as to make easier the determination of weaknesses in the preparatory work of the class. But the teacher need not have a special diagnostic test to find out many of the things she needs. As reading is the basis for proficiency in all subjects, she can test the ability of the class in both oral and silent reading very easily. This information alone will indicate whether she can expect her class to interpret the printed page with facility, and so master their textbooks without too much assistance, or whether she will have to make interpretative reading her major work. In like manner she can test for "intelligence" and other general qualities, as well as for ability in specific subject-matter.

Not only is this preliminary diagnosis of value, but it is also essential to test advancement from time to time by other than subjective tests. The test of advancement

will reveal whether the class as a whole is progressing together, or whether there are more or less well defined groups which need special treatment; groups which will justify a teacher in separating them into sections for special instruction. Where the instructor has but one grade in a room, or if, as in high school, has several sections of the same subject, she will want to arrange her pupils so that groups of like proficiency can receive similar instruction and make very nearly equal advancement. The objective test is the best school measure to be employed in making this adjustment so that pupils can move forward in groups of like attainment. The problem of sectioning, however, is not a simple one, for it is necessary to consider a number of factors other than school achievement alone. To this we shall return in a later section of this chapter.

Every set of classes in a school shows a good deal of what is known as "over-lapping." That is, when pupils in several grades are tested, it almost always is found that there are pupils in one grade who can attain the norms of one or more grades higher; and, on the other hand, there are frequently pupils who seem to be unable to meet the norms of the grade in which they are placed. Thus in a spelling test, fourth grade children are found who can spell as well as the typical seventh, or even eighth grade pupil; but fourth grade pupils are also discovered who can not meet third grade norms. The tests will tell the teacher whether enough of a given class overlap the norms of the class above, in any subject, to make it possible, other things being equal, to advance this group into the work of the higher grade either in one, or in all subjects. A number of schools

in the United States have been so organized that when these overlappings are found, rapid moving or slow moving groups may be formed to meet the situation.

TABLE I

SHOWING THE VARIABILITY IN READING ABILITIES IN GRADES VII-XII IN ONE SCHOOL SYSTEM. HAGGERTY READING EXAMINATION, SIGMA 3.

<i>Scores</i>	<i>Per Cents</i>					
	VII	VIII	IX	X	XI	XII
0-10	1	—	—	—	—	—
11-20	7	1	1	—	—	—
21-30	16	12	4	—	—	—
31-40	25	16	12	3	3	2
41-50	19	25	13	9	7	5
51-60	15	22	20	19	10	12
61-70	10	10	19	15	19	27
71-80	10	26	26	31	20	37
81-90	—	4	5	18	29	15
91-100	—	—	—	5	12	2
<i>Median</i>	41	47	62	71	76	75
<i>Range</i>	76	69	69	62	62	56
<i>Number</i> <i>of</i> <i>Cases</i>	136	113	99	67	61	41

This, of course, means that promotion by grade is subordinated to promotion by subject.

For diagnosis of individual pupils.—Closely connected with the use of the test for diagnosis of the class, is its use for determination of the difficulties and peculiarities of each pupil. While in general, individual differences are not so marked as to preclude efficient class instruction, yet the more that is known about each child's weaknesses and strong points as well, the better

success will the instructor have in handling the group. The test is to be studied especially in the light of each pupil's individual attainments and points of difficulty. This study may be the means of clearing up entirely unsuspected troubles which otherwise would have continued to hamper the child and have prevented proper advancement. As previously indicated, under our discussion of the criteria of a good test, it is necessary to know not only that a pupil is failing in a certain subject or aspects thereof, but we should know the reasons and his peculiar difficulties, as far as it is possible to discover them. In addition, the tests will disclose what pupils no longer need drill and training on a given aspect of a subject. The detection of such weakness and strength is one of the important functions of the standardized tests.

Especially valuable is this sort of individual diagnosis in the light of the various sorts of drill sheets and practice pads which have been designed in various subjects so that each pupil may drill upon his own difficulties independently of the class. The Courtis and Studebaker practice pads in arithmetic and the Courtis practice exercises in handwriting are examples of this sort of drill opportunity which carries each pupil at his own gait, without affecting the rest of the class.

Those cases where pupils are found to overlap other classes to a marked extent, so that the treatment must be individual rather than group, call for special treatment. In such cases pupils who are far beyond the norm of the class in a special subject as arithmetic may be placed with the advanced class in that subject for recitation work, keeping with the class in other subjects,

but still gaining time because of the special opportunities offered in the subjects in which they are particularly strong. Other pupils, markedly deficient in a subject, may be given the proper drill and training by being placed in a lower class where the work will be more nearly commensurate with their ability and equipment.²

Of course, a flexible and adjustable plan of this sort entails difficulties in organization and requires facilities which may sometimes prove to be great obstacles to a satisfactory program. But the results justify the effort; and consequently these special arrangements are becoming more and more common. In fact, many teachers now feel that the test is serving its greatest purpose when used to discover and encourage special proficiency, or to discover a special deficiency which may yield to special treatment.

For setting standards of achievement.—In Chapter II, it was pointed out that norms have been developed not only for the achievement to be expected of an average child in a grade, but also that norms have been worked out for many tests as standards of achievement to be aimed at in connection with the work of a year or semester. More than any other means, a definite goal to be striven for by the entire class or by the single pupil is a great stimulus to both teacher and pupil. The interest and enthusiasm of a whole class can be aroused in an attempt to attain a standard which, by the experience of other classes in the same or similar com-

² The following are some of the *plans* which have been devised to adjust school work to varying abilities of pupils: the Winnetka plan, the Batavia plan, the Portland plan, the Oakland plan, the Baltimore plan, and the Gary plan.

munities, has been demonstrated to be within the bounds of high grade work. To be sure, the diagnosis of the class must have shown beforehand that these standards may reasonably be expected, for it is bad to work for a goal which seems, and for that very reason may well be, impossible of attainment for a particular group.

It should be noted, of course, that whether standards are to be set for a class as a whole, or for small groups within the class, or for each individual, will depend very much upon the similarity or dissimilarity in the mental make-up and school training of the class. If the class is fairly uniform in ability and equipment, the standards may be more or less uniform. But if there is a wide variation in ability, the standard of one may be altogether too difficult or too easy of attainment for another. It is frequently wise, therefore, to permit each pupil to know his own status at any given time, and to keep before him the standard for which he should strive. Tests thus become potent instruments in the application of a sound psychological principle: namely, learning with knowledge of results. Beating one's own record is an absorbing occupation in almost any sort of activity, and this idea can be capitalized in getting children to work against their own records in school subjects. As a practical matter, also, the pupil's knowledge of his attainments and of standards goes far to reconcile parent and child to a rating or mark which might otherwise be regarded as prejudiced.

For promotion.—Standardized achievement tests are always valuable in *supplementing* the usual criteria of promotion or non-promotion; that is, teachers' judgments and school marks. As previously stated, their ob-

jectivity, reliability, and norms make the tests important factors in the determination of achievement. They therefore, become significant in the elimination of uncertainty in the case of a doubtful pupil. However, there should always be a combination of factors when promotion or non-promotion is being considered. Teachers' judgments and the marks for daily work are by no means to be ignored. But though every teacher try to make her ratings as objective and unbiased as possible, we know that individual judgments do not have the validity of objectively determined standards. Therefore, in most cases, and in backward or doubtful cases in particular, the additional information made available by educational tests should be taken into account. In the minds of some, the objective test is the most important criterion in determining fitness or unfitness for advancement. Not only is it valuable because of its objectivity and its norms, but it has the additional merit of being clear and understandable to pupil and parent who might otherwise see no such well defined explanation for the failure or marked success of the child.

Uses of Intelligence Tests

For class diagnosis.—After the teacher has given the achievement tests and has found how her class stands in relation to the norms in any given subject, she is in danger of making false assumptions respecting the merits or demerits of the pupils' accomplishments unless she has further information with respect to the general mental ability of the class. If the work of the class falls below the norm, the teacher may feel that her

instruction has been a failure. This conclusion might be altogether unwarranted, for the class may actually be of a low grade of mental ability, and, therefore, unable to reach the norms expected of a class of average ability. In fact, her teaching may very well have been above the ordinary to have enabled her relatively dull pupils to reach as high a level as they did. On the other hand, another teacher may plume herself on an excellent piece of teaching, whereas the very brilliancy of the group, under really strong instruction, should have carried it to a much higher level than that actually reached.³ Thus, in the one case, the apparently inferior results were not inferior at all, inasmuch as they indicated a high grade of teaching and achievement, when the actual learning ability of the pupils is taken into account; while in the second case the actual achievement is inferior to what it should have been for pupils of superior mental ability.

It is evident, therefore, that there is need for some means of determining with reasonable accuracy the actual mental ability of a class. The intelligence test meets this need. We have already mentioned some of the

³ The significance of the intelligence factor in a class was recently brought out in a study designed to evaluate a certain highly advertised method of teaching pupils individually instead of in classes, with special emphasis upon freedom of choice of work by the pupil. The pupils were tested by standard tests in the various school subjects and found to be up to or above normal standards in all. But the children had also been given intelligence tests, individually. The average intelligence quotient for the class was found to be about 120! This denotes "superior intelligence" and means that the children at that time averaged about two years beyond their actual ages in intelligence. Similar tests were made in two successive years with similar results. In the light of these facts the pupils were accomplishing much less than should have been expected of such bright children, especially as the teachers were above the average in ability. The method of teaching, therefore, was not justified in the light of the results.

intelligence tests suitable for use with various groups of pupils. By administering several of these, it is possible to determine with adequate accuracy whether any class, as a group, is up to the normal expectation in ability to perform school work.

It is true that the technical difficulties both in giving and in interpreting intelligence tests are greater than for achievement tests. But the *group* tests are so devised that they may be given by the class-room teacher; for, as will be shown in Chapter XVII, if proper precautions are taken and if directions are closely followed, the teacher's results may be regarded as quite satisfactory.⁴ The teacher may, by comparing the results of several of the group tests, make a diagnosis of the ability of the class which will go far in enabling her to determine what degree of proficiency she should reasonably expect of it in connection with the regular school work. This knowledge will enable her to plan her work more intelligently, and to prepare for overcoming difficulties which otherwise might have been unsuspected, or delayed in making their appearance, for here as in other affairs, "forewarned is forearmed."

Where intelligence testing is carried on by the supervisory force, or by special persons, time and effort are saved the teacher, for then she has only to call upon the proper authorities for information which she might otherwise have to find for herself. Strictly speaking, while achievement testing may be regarded as essentially a part of the work of the class-room teacher, in-

⁴ Some training is necessary for administering group tests, though it is brief. But only a trained psychologist should administer the individual test.

telligence testing is rather the work of the specialist. This is so especially because of the difficulties of interpreting results at times, and because of the insistent need for uniformity of procedure in administering the intelligence test.

For individual diagnosis.—The intelligence test is especially valuable to the class-room teacher in assisting to solve the problem of the proper treatment of the child who is out of the ordinary. He may be unusually bright, or dull, or mischievous, or troublesome, or in some other particular atypical. The instructor wishes to know of this child whether his general ability is as indicated by his typical responses; whether the judgments of former teachers, and, perhaps the child's own parents, are correct. For this, the intelligence test gives information not to be obtained in any other way. The result may indicate that the pupil has ability hitherto unsuspected; or that his supposed brightness is but a superficial pertness covering an actually dull intellect; or that the pupil's bad behavior comes from not keeping busy a really brilliant mind, so that his mental activity finds an illegitimate outlet in mischief and disorder, because he is bored by the comparative simplicity of the work which is commensurate only with the ability of the "average" pupil of the class.

Again, when given to an entire class, the intelligence test frequently uncovers a child of real brilliancy who has been content to go with the group without in any way showing his real ability.⁵ Such unsuspected "finds"

⁵ An interesting illustration of this fact recently came to the writer's attention. Ann was a pupil in the fourth grade of a private school, this grade corresponding to her chronological age. She was transferred to the public schools and was given the Terman achievement tests and an intel-

would never have been known, had it not been for the intelligence test.

There is overlapping in mental ability from grade to grade, just as there is in achievement, as previously described. There are children in the fourth and fifth grades who have reached the intelligence level of the normal eighth grade pupil; just as there are fourth grade pupils who are on the same level as some in the second or third grade. In fact, since intelligence tests have been applied to school problems and have been made the basis of some experimental work, it has been disclosed that in many instances there are about 25 per cent in a class who are accelerated for their *mentality*, and about 25 per cent who are retarded for their mentality. That is, in the former case, the pupils are *too far* advanced, judged by mental ability as shown by the intelligence tests; whereas in the latter case the pupils are *not far enough* advanced for their mental ability. It appears, then, that in these instances only the middle 50 per cent seem to be properly placed in the grade.

When individuals are discovered whose mentality warrants their being well in advance of their places in school, in most cases there should be a readjustment

ligence test. Her intelligence quotient was found to be 161, and her achievement test scores were above the norms for the sixth grade! The girl was placed in the sixth grade of the public school with a short period of coaching on the omitted work of the grade missed. At the end of the second month in the sixth grade she is doing better than average work for that grade.

The opposite condition existed in the case of a boy in the same system who was one of several who were demoted one grade on account of low intelligence for the grade involved. The parents of the lad objected strenuously to his demotion, but a few weeks later the boy himself bore this voluntary testimony to the principal: "I am sure glad I got sent back. You know I just couldn't understand any of that eighth grade work. Now I am getting along real well."

of work to their abilities, either by advancing them to a grade where their intelligence is given a real test, or by placing them in rapidly moving classes, so that they may make progress according to ability rather than by some fixed promotion period.⁶

The dull pupil also will be better understood in the light of the intelligence test when considered together with the results of the achievement tests. The pupil's difficulties may be understood more particularly; and what strong points he has will likely come into relief, so that the result may be an entire redirection of his instruction. In any case the intelligence test will assist both in explaining difficult cases and in revealing unsuspected strength, and perhaps weakness, on the part of apparently normal, well-disposed pupils.

Of course, too much importance must not be attached to the intelligence test alone, to the exclusion of everything else. This point has been stressed very definitely during the past few years; but, quite without justification, some persons have interpreted this caution against complete and undue dependence on tests of intelligence to mean that the tests were of no value. No such interpretation is warranted. What is meant is that the tests are not to supersede the results of class-room achievement, but to supplement them; that the tests are not to

⁶ Frequent objection is made to a plan of segregation on the ground that the placing of the slow pupils by themselves causes them to lose initiative, and removes the example of the brighter pupils. But this does not follow if the teachers are careful not to indicate that any stigma is to attach to the slow class. In fact the pupil finds that he is more at home in not being held up to comparison with the brighter children, and that he may well have a better chance to show real progress if he is competing with his own kind. Some one has said that it is always possible to get up a fat man's race, and the competitors take just as much interest as do the spectators. They are quite willing to compete in a race in which each feels that he has a fair chance to win. The parallel is a good one.

be used in place of all other information about a child, but rather in the light of such information; that the child is not to be placed in school or in other relations of life simply on the basis of his intelligence quotient (IQ), but rather that his IQ is to serve as a means of interpreting otherwise undiscoverable factors of his personality.

The combination of achievement and intelligence test results.—In order that undue stress may not be placed upon the child's proficiency, or lack of it, in any one sort of test, the well-informed teacher and administrator will attempt to rate the pupil in the light of his various attainments, or indices. A very important one among these is the child's "mental age"; that is, the degree of mental ability (as measured by the tests) which is possessed by the average child of the corresponding chronological age. It is "an index of absolute mental level which indicates the level of development which a child has reached at a given time."⁷ For example, if, on a certain mental test a child earns the score of the average ten-year-old, he is said to have a "mental age" of ten; if he earns the score of the average eight-year-old, he is said to have a "mental age" (abbreviated MA) of eight, etc. It is impossible here to go into the assumptions, bases, and limitations of the MA;⁸ but suffice it to say that as an index of ability it has high validity and marked significance for the teacher and in school problems.

A child of any given chronological age (CA) may score above or below the norm for his age, and he is,

⁷ L. M. Terman, *The Intelligence of School Children*, Chapter I.

⁸ F. N. Freeman, *Mental Tests*, Chapters IV and XVIII.

accordingly, rated as being superior or inferior, in varying degrees, depending upon the extent to which he exceeds or falls short of the norm. Suppose that a test shows a child of nine years and three months (chronological age) to test as high as the "normal" child of ten years and six months (chronological age); the former child is then said to have a "mental age" (MA) of ten years and six months.

Now it is not enough to know only that a child has a certain MA; it is important that we know how long it took that child to reach the level of intelligence indicated by his MA; that is, we must know the chronological age as well. The relationship between the MA and the CA is known as the intelligence quotient (IQ), which expresses the ratio of the mental age to the chronological age.⁹ In the instance cited above it would be:¹⁰

$$IQ = \frac{126 \text{ (MA)}}{111 \text{ (CA)}} = 113.5 \text{ (or 114)}$$

It is clear that inasmuch as there are individual differences in the rate and level of mental development, just as in physical development, it is necessary to express the relationship between these two factors of rate and level. The "mental age" may be regarded as expressing the level of mental development at the time of measurement, while the IQ expresses the rate, which is interpreted in terms of brightness or dullness. Thus, three children may all have a "mental age" of 10; yet one

⁹ For an account of the significance of the IQ, see L. M. Terman, *The Measurement of Intelligence*, Chapter VI.

¹⁰ Years are reduced to months. Decimal points are disregarded in the quotient.

is, let us say, 8 years of age (CA), another 10, and the third 12. We should be under a serious misapprehension if we regarded these three individuals as being equally intelligent, for it has taken them 8, 10, and 12 years respectively to reach *the same level*. But the IQ will show this disparity for the intelligence quotients will be as follows:

$$\frac{120 \text{ (MA)}}{96 \text{ (CA)}} = 125; \quad \frac{120}{120} = 100; \quad \frac{120}{144} = 83$$

The first, according to the usual classification, would be regarded as "very superior," the second as "normal" or "average," while the third would be designated as "dull." Thus the value of the MA is decidedly enhanced by its association with the IQ, and the two, taken together, supply the teacher with an item of information which is of extreme importance in the judgment and treatment of a pupil.

In her judgment of a pupil, however, the teacher has available other measures and indices. By means of appropriate tests, she has measured his ability in reading, arithmetic, spelling, history, etc. These standardized tests supply the teacher with norms for various ages or, much more frequently, for grades. Using the pupil's scores and the given norms, she is able to state his "achievement age" for the several subjects; and employing the same technique as in the case of the IQ, she can derive a ratio between the "achievement age" in the subject and the child's chronological age. Although norms for achievement tests are usually given for school grades, rather than for ages in years and months, yet by using the ages ordinarily assigned as the

“normal” or average for the grades, a useful and fairly accurate index may be worked out. Thus, when a seventh grade norm is given for an arithmetic test, that same norm may be spoken of as the norm for the average age of the seventh grade pupil. If we should assume, as an example, that the “normal” age for the seventh grade is thirteen years, then a child reaching the subject norm of the seventh grade may be said to have an “achievement age” of 13 in that subject.

Based on the results of investigations by Ayers,¹¹ Terman,¹² and Kelley, McCall¹³ has presented a table which gives the average age of the first grade child as 80 months at entrance into school, and adds 13 months for each succeeding grade, inasmuch as the investigators named above seem to agree that the average increase in age from grade to grade is between 12.5 and 13 months. On this basis, the average age of a seventh grade child, in September, is 158 months; or, if as usually happens the norm is for the month of May instead of September, the age would be 167 months.

It will be observed, of course, that by using standard tests in various subjects it is possible to secure “achievement ages” in these subjects. Thus a child may have a “reading age” of 9, an “arithmetic age” of 10, and a “spelling age” of 9. But it is desirable to have an expression or term which will combine these separate “ages” and show in general what the child’s school attainment is. For that purpose we have the “educa-

¹¹ Leonard P. Ayers, “The Relation Between Entering Age and Subsequent Progress Among School Children,” Bulletin No. 112, Russell Sage Foundation, New York City.

¹² L. M. Terman, *The Intelligence of School Children*, p. 94.

¹³ W. A. McCall, *How to Measure in Education*, pp. 34 ff.

tional age" (EA), which is the average of the "achievement ages" for the individual school subjects. In the case of the child mentioned above, the educational age would be

$$\frac{9+10+9}{3} = 9.3 \text{ years (or 9 years, 4 months)}$$

The EA is frequently a very useful index, for it indicates the *average* of a pupil's school achievement. But therein also lies its *weakness*; for, being an average, it may conceal special proficiency or special disability. Consider, for example, the pupil who has the following record:

reading age	10
arithmetic age	8
spelling age	11
language age	11

Averaging these "ages" we find that the EA is 10. Now if the pupil is approximately ten years in chronological age we should say, on the basis of the EA, that he is doing school work which is normal for his years. But we should be overlooking the fact that this same pupil is seriously retarded in arithmetic ability. It is likewise apparent that the EA might conceal a special *ability*. Though the EA is important in a general way, and especially for the purpose of studying *groups*, it is imperative that we go behind this index to examine the items which determine it, if we get a true and significant picture of an *individual*.

It was stated above that an "achievement age" could be used with the chronological age in the same way as the mental age, but to determine, of course, an "educa-

tional quotient" for the school subject.¹⁴ Though this can be done, and is done, it is customary to determine an "achievement quotient" in a somewhat different manner. The achievement quotient (also known as *accomplishment quotient*) may be found thus:

$$AQ = \frac{EA}{MA}$$

Let it be noted that the AQ is the ratio of the educational age to the *mental age*, and not to the chronological age. The mental age is used as the denominator because it is assumed to be a more reliable index of ability to do school work than is the chronological age. There are a number of statistical and psychological objections to the AQ; but with respect to its usefulness as an empirical and meaningful device in the school there is little question.

How shall the AQ be interpreted? We may say that it gives us a good indication of whether a pupil is achieving as much as might reasonably be expected of him, judging from his mental ability (MA) as measured by the intelligence test. In general, pupils having a low AQ probably need to be stimulated or are suffering from unfavorable conditions, physical, emotional, or environmental. Such pupils should be studied individually to determine the source of difficulty and to apply a remedy where possible.¹⁵

As stated in an earlier section of this chapter, the

¹⁴ Educational Quotient (EQ) = $\frac{\text{Achievement Age}}{\text{Chronological Age}}$

¹⁵ For a theoretical criticism of the AQ see F. N. Freeman, *Mental Tests*, pp. 285 ff.

evaluation of the achievement of any class of pupils, and the consequent judgment of the teacher's efficiency, cannot rest upon the results of the subject-matter tests alone. We must know the nature of the *pupil material* with which the teacher is working; in other words, we must know the average "mental age" and the variability¹⁶ or spread of "mental ages" in the class. Knowing the "mental ages" of the pupils, it is possible to derive AQs for each class and to determine thereby whether the class as a whole is achieving what might reasonably be expected of it, in the light of its general mental caliber. Where age norms are not available for an achievement test, it is possible to form rather reliable judgments of the achievement of a class by comparing its intellectual status with that of a "normal" class, and then doing likewise with the results of the subject-matter test. This method lacks the objectivity which attaches to numerical indices, but it is significantly reliable and distinctly superior to judgments based on personal opinion, which generally ignore the relationships between achievement and mental caliber.

Summary.—At this point let us summarize this discussion by noting the important indices which have thus far been dealt with: (1) the *norm* for an age or grade group; (2) the *mental age*, indicating the level of mental development; (3) the *intelligence quotient*, indicating the degree of brightness, or rate of mental development; (4) the *achievement age* for a single subject; (5) the *educational age*, being the average of the achievement ages; (6) the *accomplishment* (or *achievement*) *quotient*.

¹⁶ Variability is considered in Chapter XVIII.

Factors to be considered in the classification of pupils.—Without going into great detail, we may point out the important factors which must be considered together in classifying pupils. The following are significant in cases where pupils are being placed in one of several classes, or where they are being sectioned within the same class. (1) *The "normal" age for the grade*; and, therefore, the corresponding "mental age" for that grade must be known. (2) *The pupil's MA*. Is it approximately "normal" or is it above or below? Is it sufficiently retarded or advanced to warrant demotion or promotion? (3) *The pupil's IQ*. Is it such as to indicate slow or rapid progress? In other words, is it at "normal," above, or below? It is impossible to say that a child with such and such an IQ shall be placed in this or that group. Once having determined that the MA qualifies the child for a certain grade, the division into groups on the basis of IQs will have to depend upon the facilities of the school in question. (4) *The pupil's "achievement ages" and his "educational age."* Is the child's actual school learning such as to qualify him for placement in a certain grade? A very bright child whose MA places him several grades above that in which he finds himself should not necessarily be moved up to the higher grade at once, for it is conceivable that he may lack the specific school equipment which will be necessary for successful performance in that grade. The pupil's program should be so planned that in due time he will be doing the level of school work for which his mentality apparently qualifies him. This does not necessarily mean, however, immediate "skipping" of grades.

The four considerations here noted are most essential and must be evaluated as aspects of a total, in the case of any pupil. It is neither possible nor wise to attempt to prescribe a formula for use in classifying and sectioning groups of pupils. Each child is a problem in himself, to be studied and accorded the sort of treatment which best suits his own needs. In addition to the factors already mentioned, it is sometimes necessary to take into account such items as the health and physical development of the child, his social behavior, his extra-school activities, etc. In some instances these factors will make it seem inadvisable to accelerate a child; in others they will have no rôle to play.

The tests which we have been discussing—of both school achievement and intelligence—are not perfect; but their wise use will contribute much to a solution of problems of the school and class-room. They place school procedures on a more nearly accurate and objective basis.

SELECTED REFERENCES

Ayers, L. P., "The Relation Between Entering Age and Subsequent Progress Among School Children" (Bulletin No. 112, Russell Sage Foundation, New York City).

Chapman, J. C., "The Unreliability of the Difference Between Intelligence and the Educational Rating" (*Journal of Educational Psychology*, Vol. 14, pp. 103-108, 1923.)

Dearborn, W. F., *Intelligence Tests* (Boston, Houghton Mifflin Company, 1928).

Dickson, V. E., *Mental Tests and the Class-room Teacher* (Yonkers-on-Hudson, N. Y., World Book Company, 1924).

Franzen, R. H., "The Conservation of Talent," Chapter IV. *Intelligence Tests and School Reorganization*, Terman, et. al. (Yonkers-on-Hudson, World Book Company, 1922).

Franzen, R. H., "The Accomplishment Quotient" *Teachers College Record*, Nov. 1920.

Freeman, Frank N., *Mental Tests* (Boston, Houghton Mifflin Company, 1926).

Hines, H. C., "What Los Angeles is Doing With the Results of Testing" (*Journal of Educational Research*, Vol. 5, No. 1, Jan., 1922).

Keener, E. E., "The Use of Measurements in a Small City School System" (*Journal of Educational Research*, Vol. 3, No. 3., March, 1921).

McCall, W. A., *How to Measure in Education* (New York, The Macmillan Company, 1922).

Odell, C. W., *Traditional Examinations and New Type Tests* (New York, The Century Co., 1928).

Pintner, R., *Intelligence Testing* (Henry Holt and Company, 1923).

Stebbins and Pechstein, "Quotients, I, E, and A" (*Journal of Educational Psychology*, Oct. 1922).

Terman, L. M., *The Intelligence of School Children* (Boston, Houghton Mifflin Company, 1919).

Terman, L. M., *The Measurement of Intelligence* (Boston, Houghton Mifflin Company, 1916).

Toops, H. A., and Symonds, T. W., "What Shall We Expect of the A. Q.?" (*Journal of Educational Psychology*, Vol. 13, pp. 513-528, 1922, and Vol. 14, pp. 27-38, 1923).

Trabue, M. R., "Some Pitfalls in the Administrative Use of Tests" (*Journal of Educational Research*, Vol. 6, No. 2, Sept., 1922).

CHAPTER IV

REQUISITES FOR GIVING OBJECTIVE TESTS

The technique of objective tests.—The earliest tests which were made public were so constructed that the typical class-room teacher would have had difficulty in using them. In fact, it was the early idea that only a trained psychologist should attempt to give or evaluate such measures. But as time has gone on, it has been found possible so to modify the nature of many of the tests and to give such careful and adequate directions for their use, that almost any intelligent person with experience in handling children can secure satisfactory results. In this chapter will be set forth some of the principal factors which enter into the use of tests, if they are to be accurate measures. After studying these various requisites, each teacher will have a very good idea of her own adaptability to the work, and will know what preparatory training may be necessary.

There are three steps to be considered in testing: (a) administering the test; (b) the scoring of the papers; (c) the interpretation of the scores.

ADMINISTERING THE TESTS.—The requisites for giving tests properly come under four principal heads: (1) preparation and training; (2) adherence to directions and procedure; (3) pupil control; (4) proper environment.

Preparation necessary.—The person giving tests must be temperamentally fitted to follow directions to the letter. That teacher who is so fond of showing "individuality" that she can never do things as others do them is temperamentally unfitted for testing. For if there be any deviation from the instructions which are to govern the giving of the test, there will be deviations in the results, and the scores will be worthless. Testing is scientific work and must be done by scientific methods; one prime quality of scientific method is accuracy. This refers to the language in which test directions are given; to time limits which are imposed; to preliminary arrangement and preparation of pupils; to voice modulations and quality. The substitution of a single word in giving instructions or asking test questions may alter the entire situation. The inflection of the voice may give a pupil the hint to an answer or a cue to a situation or response in a way not at all intended. The addition of a quarter of a minute to the time allotment may have no appreciable effect, but it may change an entire class rating and render comparison with stated norms meaningless.

Therefore, the teacher must follow directions to the letter. If unusual situations arise, she is to give the test as instructed; and in evaluating it, if there be doubt about the validity of the results, she should submit them with the attendant circumstances to a trained expert in the school system for an opinion; or if there be no one of the sort available, to the author of the tests, or to a person qualified in educational psychology. Time elements must be observed to the second. The voice must be kept at as even a pitch as possible;

voice inflections must not reveal any emotional disturbance in the teacher, which would in any way invalidate the intent of the directions.

All of this means a definite preparation for giving the test; a careful study of the directions; a rehearsal orally of the entire test before appearing before the class with it. If possible, the teacher should practise by giving the test to other adults, and should also take it herself. Her preparation should result in an absolute familiarity with the test and with the possible contingencies which might arise in connection with it.

In like manner a study should be made of the best preparation, arrangement, and seating of the pupils who are to be tested. A very proper method is to arrange the group as for the proposed test, and then give the group a brief subjective test along somewhat similar lines to accustom them to the sort of thing which will come in the actual test, carefully refraining from anything which would be practice of the actual situations involved in the standardized measure, except for the general arrangement and attitude of the children. Many tests are now constructed with preliminary exercises intended to give the children an idea of the sort of thing to be required of them, and used as preliminary to the main test to insure an absolute understanding of the requirement of the test. These are called "warming up" exercises, or "shock absorbers." Where these are available, they will be found to be of great assistance in preparation for the real test. Most measures are constructed with the thought that they will be given to regular class groups; so in many cases this preliminary arrangement is unnecessary. But the

teacher should satisfy herself on that ground, in order that she may not find an unfamiliar, extraneous factor introduced which might have been foreseen and avoided.

Though not always possible, it is highly desirable that the teacher have an opportunity to observe and to be observed by an experienced examiner. By so doing she may more readily learn the necessary qualities for successful and reliable testing. If such observations are not available to the teacher, she may undertake to train herself through careful adherence to test directions and through practice. Her first results will not be so reliable as they might otherwise be; but she will discover what difficulties arise and how to remedy them. This means, of course, that under these conditions the early test results must not be over-evaluated.

Strict adherence to directions necessary.—The teacher who is not self-controlled, who is lacking in poise, is likely to fail in giving the tests. Undue emotionalism of any sort is incompatible with scientific accuracy. The teacher who allows impulse to rule her actions, who is overly sympathetic with the pupil who is having difficulty with his test, who can not resist giving hints, or "little helps," who is thinking more of the showing made by the pupil, than of the value of the test itself, is not to be trusted with such scientific devices. The teacher who yields to such tendencies can not compensate for this laxity by making allowances in any other direction, without making a bad matter worse. A certain high school principal, very capable as an administrator, but without adequate knowledge of tests, a short time ago attempted to give a certain test,

but decided that the time limit was too short, and so allowed double the time required. When told that this invalidated the test, he said: "I can fix that all right; I shall just divide the scores by two!" A better example could not be given of failure to appreciate the scientific attitude and the meaning of scientific precision. The teacher "who likes to do things her own way" and so modifies her manner and attitude as to defeat the purpose of the test, either by actually failing to follow directions, as has been already suggested, or by interpolating remarks which she thinks will be of assistance to the pupils, is also to be forbidden the use of the test. On the other hand, that teacher who in general has developed a feeling of antagonism or of fear in her classes, who is totally unsympathetic, harsh, or bad tempered, will find that the class will not respond properly to the test.

The teacher who is even-tempered and self-controlled in every way possesses the emotional requisites for becoming the ideal examiner. Absolute honesty of purpose and attitude is, of course, presupposed. Naturally, any other type of person is not qualified to teach or to examine pupils.

Pupil control necessary.—Test results mean nothing unless the class or group is under definite control. Order is as essential to testing as to every successful school exercise. The teacher who does not know how to secure and maintain good order cannot succeed in testing work. The group must be under such control that it is in sympathy with the purpose of the teacher, whether during an ordinary recitation or during an important test. Pupils work best in tests, as in every

other sort of work, when they are in a frame of mind as nearly natural, as little flurried or agitated, as possible. The teacher who can maintain sympathetic control will have such an attitude in the group, and will find that the result of the test comes as near being accurate and representative as can be desired.

Further, the examiner must be cautioned not to give a test when the group is agitated or "upset" by any untoward circumstances. Lapses from ordinary class conditions may decrease the control which is so necessary; tests should, therefore, not be scheduled when there are likely to be any unusual conditions. Thus the day before a holiday or before a big athletic contest, the morning preceding a school outing, or any other such event, should be avoided as a time for testing, because of the poor conditions of control.

There are also certain hours of the day which are less desirable than others. In general it may be said that periods immediately after the opening of the session and periods immediately before the closing of the session are least desirable. This applies to both morning and afternoon. It is probably most advisable to administer tests in the morning, particularly so in the case of younger children.

Good environmental conditions necessary.—This leads to the fourth factor, namely, the environmental conditions. The test can be best given when there is absolute quiet. If the ordinary class-room is so situated as to be affected by the noise from the street or nearby industry, for purposes of testing an exchange should be made for the period with some other class-room better suited to the purpose. Except for very small

children, the change of environment of the strange class-room is not likely to be so serious a factor as the presence of noise. The test should not be given when there is a recess for a part of the school, making a noisy playground situation which may inhibit the proper attention of the pupils to the test.

Interruptions must be avoided. So the test should not be held at a time when there is any likelihood of an interruption from visitors or from a fire drill, or when there is any possibility that the time ordinarily allotted may be curtailed, or danger of any other type of distraction which would invalidate results, either for the group or for individuals. The tests are so important that they should not be exposed to the danger of modification by any such preventable situations.

Any factors which might induce undue strain are to be avoided. Presence of a principal or superintendent may produce such a situation. In such cases, if it can be managed, these supervisory officers should be persuaded to remain out of the room. Pupils may be actually frightened by unusual stress being placed upon the importance of the test and thereby unfitted for their best efforts. In general the less that is said to the groups about the importance of the tests, the less strain is likely to result. Certainly the attitude of some teachers in advertising that they are to give tests is not conducive to the most satisfactory results, or the best environmental influence.

THE SCORING OF THE PAPERS.—It will be remembered that one of the great objections to the subjective type of test or examination is the variability of marks assigned to the same papers by different teachers, or to

the same paper by the same teacher on re-scoring. One of the great values of the objective test is that it removes this difficulty for the most part. But this result will not hold unless as much attention is paid to following directions in scoring as in giving the paper. The teacher will find it possible to obtain specific instructions as to methods of marking the tests, and she must follow these instructions without deviation, or the result will be invalidated. A teacher may not altogether agree with the principle of marking used by the author of the test, but she may not for this reason change the method of procedure. The standardization of a test means that absolutely uniform methods have been used both in giving and in scoring, and any departure from these methods will make the test worthless. For these reasons the instructor must remember that all elements of subjectivity must be eliminated.

One great advantage of most tests and scales is the ease with which they may be corrected. Not only are instructions given for the method to be followed, but devices are suggested for reducing the work to a mechanical system, in all cases where no element of judgment is involved, as is the case in practically all of the intelligence tests, and in many of the achievement tests and scales. The drudgery of correction is therefore much less than in the older type of examinations; so much so, in fact, that many teachers are modeling the regular subjective examinations upon the plan of the objective tests. For this reason, the teacher need not fear that the giving of the test will involve a great deal of extra time for correction.

TABULATION OF RESULTS.—After the tests are cor-

rected, the next step is to tabulate the results. Many tests are accompanied by instructions for some method of setting down the marks of an entire group or class in such a way as to be easily read and referred to. In general, such methods are the result of long experiment and will enable the teacher to use her data to best advantage. In order to interpret the tests in the light of the plan of their construction and application, the suggested methods of tabulation and arrangement should be followed, even if they involve a somewhat different type of work than has been a part of the past experience of the teacher. This refers especially to methods of constructing graphs or using other methods of charting results. Directions such as those accompanying the Courtis Practice Tests in Arithmetic very clearly tell how to make such graphs and give illustrations of them. Therefore the teacher will not find the task of making ordinary graphs a difficult one. In fact when she has once learned to read a graph intelligently, she will prefer this method of presenting results to any other.

In the final chapter we shall consider in some detail the kinds of graphs and their uses. But at present it is important to observe that ordinary graphs are especially useful in offering a pictorial representation of numerical data. To the person of even limited experience the graph will yield information which the numerical data themselves may conceal.

THE INTERPRETATION OF THE SCORES.—After the papers are marked and the scores tabulated and graphed, the teacher is then ready to make use of them. Many persons think of this matter of interpretation as one in-

volving a knowledge of complex statistics, extensive psychology, and unusual experience and judgment. Truly enough, all of these factors are involved in making and standardizing the test. But the class-room teacher has had the way prepared by the builders of tests, so that she need anticipate no great difficulty in the more general and obvious forms of analyzing and applying the results. Among the earlier forms of analysis are the determination of the class average and its comparison with the grade norm; the comparison of any given pupil's score with the class average or the grade norm; the observation of variability within the class; observation of overlapping from grade to grade; observation of improvement in the school subject as shown by an increase in the class average and in individual scores, etc.

In the descriptions of the various tests included in the following pages, their uses will be indicated and suggestions will be given as to conclusions which may be drawn from the results. When the tests themselves are obtained, it will be found that they are accompanied by data and instructions for their use and interpretation. These data and instructions will, of course, simplify the problem of what conclusions may be drawn from one's own results.

In order to make the language of interpretation entirely intelligible, a few terms commonly used in measurement and statistics must be mastered. The median, the quartile, probable error, deviation, and similar terms are used so frequently that no teacher can afford to be ignorant of them. On a proper comprehension of the more common statistical terms depends, in large

part, a proper grasp of the problems presented by test results.¹

In Chapter XVIII we shall present some illustrations of the way in which certain concrete results may be interpreted. These illustrations, better than any descriptions, will indicate the kinds and methods of analysis.

The principal idea which this part of our discussion should leave in the mind of the reader is that giving and scoring of a test is worth little in itself. The *use* to which the measure is put gives it its principal value. Many persons have been quite content to say: "I gave such and such a test to my classes last year," and have therefore felt that they should be congratulated on their progressiveness, when, in fact, they did not actually make use of one bit of data which might have been gained from the test. Therefore the teacher should look into the use which she is to make of the measure, the way in which she is to interpret the results, familiarize herself with the real purposes of the tests, and then put all this knowledge into practice, or she should not give them at all.

¹ Extended treatment of statistical devices will be found in any of the texts, such as *Educational Statistics*, by Odell (New York, The Century Co.); *Statistics in Psychology and Education*, by Garrett (New York, Longmans, Green & Co.); *Statistical Methods for Students in Education*, by Holzinger (Boston, Ginn and Company).

CHAPTER V

SPELLING

Early objective tests.—The earliest reported objective tests in education were those of the Rev. George Fisher, an English schoolmaster who, in 1864, was the author of a book giving questions and samples to permit objective numerical grading in “writing, spelling, mathematics, navigation, Scripture, knowledge, grammar, and practical science.”¹ In this country the earliest known beginning was made by J. M. Rice, who in 1897 reported an investigation in spelling at a meeting of the National Education Association in St. Louis. His report marks the first public announcement in this country of the modern practice of measuring the results of teaching. Rice selected a list of fifty words and submitted them to a number of schools in order to determine the effect of different amounts of drill on ability to spell. His list of words was not scientifically selected, nor was it standardized in the sense that he knew what score an average third or fifth grade child should make. Yet it is worthy of notice that Rice used this first objective test as a means for determining the relative effectiveness of different methods of teaching. Instead of trying to settle the problem of methods by the traditional plan of opinion and debate,

¹ Thorndike, E. L., “Educational Measurements Fifty Years Ago” (*Journal of Educational Psychology*, Vol. 4, November, 1913).

Rice offered a measuring instrument for the determination of efficiency. This idea was new to teachers and administrators, and it opened the way to a careful scientific study of the problems involved in the learning and teaching of spelling. Let us further consider what some of these problems are.

Some problems of spelling.—Spelling depends upon the formation of certain associations as a result of experience. In spelling a word, we may recall how the word looked or sounded when we spelled it, or when it was spelled for us; or more likely, motor imagery carried the hand along in the writing process without much thought about the particular letters of the word. With the child, the problem of forming correct associations for spelling is an extremely hard one, for the unphonetic character of the English language is a serious difficulty in the learning of spelling. This is demonstrated by the variety of sounds for the same letter and the variety of letters which may represent the same sound.² Silent letters further complicate the process. For these reasons it is necessary that the child learn independently almost every word he uses. Inasmuch as nearly all of his spelling is used in written work, motor learning prevails in order that the words may be learned under conditions of actual use. These facts complicate the child's problems in spelling.

The different methods for teaching spelling may be classified under two general headings. It may be taught in connection with reading, composition, and the other

² For a demonstration of this fact see "A Source of Confusion in Spelling," by Horn, E., (*Journal of Educational Research*, Vol. 19, No. 1, pp. 47-55).

school subjects by what is called the incidental method. On the other hand, it may be divorced from the other school subjects and taught by the drill method. The words may be spelled orally or written. They may be used in sentences or out of context. A further problem of importance to the teacher is the selection of words that should be taught. There are more than 350,000 words listed in the New International Dictionary, but Jones³ found only 4,532 different words used by school children in their compositions. Terman⁴ gives the number of words in the vocabulary of the "average" eight year old child as 3,600 and that of the "average" adult as 11,700. In another experiment⁵ conducted with pupils in grades 4 to 8, it was found that 4,515 words were given in response to certain "stimulus-words." Of these, 1,309 were common to all the grades tested and form a very large portion of the vocabularies in each of the grades. Just what words, therefore, are to be selected by the teacher for spelling drill? This problem must be solved in terms of experiments such as those mentioned above and like those to be described in the following pages.

The problems of spelling are not theoretical but practical problems of the class-room teacher. Stated more concretely, the teacher wants to know whether she is using the best methods in teaching spelling, whether she is devoting too much or too little time to the subject, or whether she should teach spelling along with

³ Described in a later section of this chapter.

⁴ Terman, L. M., *The Measurement of Intelligence*, p. 226.

⁵ Shambaugh, C. G. and Shambaugh, O. I., "A Core Vocabulary for Elementary School Pupils" (*Journal of Educational Research*, Vol. 19, No. 1, pp. 39-46).

the reading; or, more important still, whether she is teaching the children to spell the words they most need to know. She may be positive in her own opinion and give no further consideration to the problem. Unfortunately, the strength of her opinion may bear no close correlation with the facts in the case. If the teacher attempts an answer to the problem by testing her pupils by the traditional methods, she will select a list of words from the speller, the reader, or elsewhere and give them to the class. One child may score 90, another 65, and another 40. But what should they score? That depends upon the words. The only way to know what the pupil should make on a spelling test is to use a standard measure in which we know the norm for a child of a certain age or grade. Some of the more important of these standard tests will now be described.

THE AYRES SPELLING SCALE

Description and derivation of the scale.—This spelling scale was devised by Leonard P. Ayres. It consists of 1,000 words arranged in twenty-six columns with from two to eighty-two words in each column. The words in each column are of approximately equal difficulty, and the columns are arranged in order of difficulty with the easiest words in the first columns. The larger number of words are in the middle columns, and there are a smaller number of words in the columns toward either end of the scale.

The first problem in the construction of a spelling scale consists in the selection of the words to be used.

Ayres began by listing 368,000 words found in business letters, newspapers, and literature. From this list he selected the 1,000 words recurring most frequently—in these three sources—as representing the most commonly used words in the English language. It is interesting to note in this connection that fifty different words appeared so frequently that they made up about half of the total list of material examined.

The next problem was the arrangement of these 1,000 words into a scale. In order to determine the relative difficulty of the different words, they were submitted to 70,000 children from the second to the eighth grades in representative schools in widely separated parts of the country. On the basis of the number of times a word was misspelled the words were arranged into twenty-six lists. The words most often spelled correctly were placed in column "A" at the beginning of the scale and the most frequently misspelled words in column "Z" at the end of the scale.

Method of giving and scoring the scale.—The teacher in using the scale selects a list of words from one of the Ayres columns. Any number of words may be selected, but if reliable scores for the individual members of the class are desired, at least twenty words should be chosen. In general it is best to select words from a column in which about 75 per cent of the spellings are expected to be correct. This gives sufficient range for both the best and the poorest spellers in the class. The words may be given either in or out of context. The pupils' papers are scored in the usual way by determining the per cent of words spelled correctly.

Norms ⁶ are given at the top of the scale for each list of words. At the lower end of the scale norms are given only for the lower school grades. In the middle portion of the scale norms for as many as four or five grades are given for each list of words. At the upper end of the scale norms are given only for the upper grades.

BUCKINGHAM EXTENSION OF THE AYRES SCALE

Description and derivation of the Buckingham Extension.—B. R. Buckingham has made an extension of the Ayres list of 1,000 words by the addition of 505 words. These words were derived from the relative frequency of occurrence in a number of spelling books. They are, therefore, as Buckingham points out, not strictly an extension of the child's fundamental vocabulary. The difficulty of these words was determined and the words placed at the end of the Ayres columns. In general the additional words occur in the middle and upper end of the new scale.

Function of the scales.—The Ayres and the Buckingham-Ayres Scales are more than ordinary measuring devices. The method of selection makes the material fundamental in the teaching of spelling. In other words, the class-room teacher can well afford to drill her pupils on this list of 1,505 words instead of the usual large number of less frequently used words.

⁶ In this instance the norm is the score of the "average" child in any given grade. For example: the norm for the sixth grade on the Ayres Scale for the words in Column "S" is 73. That means that the "normal" sixth grade child should make a grade of 73 in spelling a list of words taken from this column. These norms were derived by the number of correct spellings for each word by the 70,000 children for the different school grades as described above.

A	B	C	D	E
99	98	96	94	92
	THIRD GRADE →	100	99	98
				FOURTH GRADE →
me do	and go at on	a it is she can see run	the in so now man ten bed top	he you will we an my up last not us am good little ago old bad red

SECTION OF THE BUCKINGHAM EXTENSION OF THE AYRES SPELLING SCALE

All the words in each column are of approximately equal spelling difficulty. The steps in spelling difficulty from each column to the next are approximately equal steps. The numbers at the top indicate about what per cent of correct spellings may be expected among the children of the different grades. For example, if 20 words from column H are given as a spelling test it may be expected that the average score for an entire second grade spelling them will be about 70 per cent. For a third grade it should be about 92 per cent, for a fourth grade about 98 per cent, and for a fifth grade about 100 per cent.

AB	AC	AD	AE	AF	
← THIRD GRADE					
2	1	0	← FOURTH GRADE		
6	4	2	1	0	← FIFTH GRADE
12	8	6	4	2	← SIXTH GRADE
21	16	12	8	6	← SEVENTH GRADE
34	27	21	16	12	← EIGHTH GRADE
50	42	34	27	21	← NINTH GRADE
combustible guarantee incessant lieutenant occurrence pneumonia proficiency villain	abyss cantaloupe embarrass- ment poulcie sovereign syndicate	appendicitis chauffeur hippopotamus maneuver miscellaneous penitentiary souvenir	hallelujahs inflammable rhinoceros	conscientious discernible dissension jardinier naphtha rendezvous	

This scale is not a test but a list of words from which a teacher can make a test. The words in any column are approximately equal in difficulty; and it is best, therefore, to choose all of the words for a test from a single column.

Twenty words are enough to secure a reasonably reliable measure of the spelling ability of a class; but for such a measure of the ability of an individual 50 to 100 words will be required. Thus, owing to the fewness of the more difficult words, it may be necessary in testing upper grades to use words from more than one column. In such cases the differences in difficulty must be recognized.

In order that the words may be difficult enough really to measure spelling ability, they should be selected from columns for which the standard per cent of correct spellings is close to 50—say between 50 and 66.

The most appropriate measure of spelling ability is secured when the words are dictated in sentences at approximately the standard rate of hand-writing for the grade in question, no test word occurring at the end of a sentence. The placement of words on this scale, however, is on the basis of returns from column dictation. Children spell more accurately when they write words in columns than they do when they write them in sentences. If, therefore, words are dictated in sentences, as suggested, results may be expected to be somewhat lower than the scale indicates. It was found that the words in each column from A through G fell three columns to the right when dictated in sentences (untimed); that those in columns H through Q fell two columns to the right, and that those in columns R through V fell one column to the right. No difference, due to dictation in sentences rather than in columns, appeared to exist for words harder than those in column V.

The 505 words added to the Ayres Scale by Buckingham are printed in italics. They were not chosen, as Ayres' words were, according to frequency in use in written discourse, but rather according to agreements among spelling books. They are not, therefore, offered as constituting a fundamental vocabulary in the same sense as do the original 1,000 words selected by Ayres. The original words of the Ayres Scale are printed in Roman.

occurring in the ordinary spelling book. As Ayres states, the children should be so thoroughly drilled on the words that the scale would no longer be a measure of spelling ability.

While the list of words, because of the subject-matter from which they were chosen, may not be representative of the written vocabulary of children,⁷ it is much more so than usual spelling lists. The scale has a further advantage in its simplicity in giving and scoring. It is one of the relatively few standard tests that the average grade teacher can administer without special training in the technique of giving tests.

MONROE'S TIMED SPELLING TESTS

Description and derivation of the tests.—Monroe selected the words for his test from the Ayres spelling list and placed the words in sentences. There are three tests. Test I is composed of 22 sentences for use with the third grade and 22 sentences for use with the fourth grade. Each group of sentences contains fifty words from column "M" of the Ayres list. Test II is similar, the words being taken from column "Q" of the Ayres list. One set of sentences is for use with the fifth grade and the other set for use with the sixth grade. Test III contains words from columns "S", "T", and "U" of the Ayres list incorporated in two groups of sentences, one for use with the seventh grade and the other for the eighth grade and the high school.

Method of giving and scoring the tests.—These are

⁷ For a more representative list see reference to Thorndike's *The Teachers' Word Book* in a later part of this chapter. Also Shambaugh and Shambaugh, *op. cit.*

timed tests in that the sentences are to be dictated by the tester at a certain rate. This rate is 10 per cent slower than the average of handwriting as determined by Freeman for each of the school grades. The sentences must be read very distinctly with no repetitions. If a child cannot complete the sentence within the time, he is to write as much as he can and then go to the next sentence. He is told before the test begins that if there are any words that he cannot spell he is to omit them.

The words taken from the Ayres list are italicized in the Monroe tests and only the italicized words are counted in scoring the papers. Since there are fifty words in each list, two points credit is given for each word spelled correctly. Norms for the different grades are given as follows:

NORMS FOR MONROE'S TIMED SPELLING TESTS

<i>School Grade . . .</i>	3	4	5	6	7	8	9	10	11	12
<i>Monroe Norm (%):</i>	56	78	66	80	70	84	86	90	94	96

Thus the "average" third grade child should make a score of 56 per cent on the test for the third grade; and the "average" seventh grade child should make 70 per cent in the test for the seventh grade.

Function of the tests.—These tests have the advantage that the material is presented in a context which is the most natural way of spelling. Very seldom is a person called upon to spell a word except in writing. For this reason these tests are superior to most of the other spelling tests in this respect, and the teacher wishing to determine the ability of her class

to spell words *in sentences* will do well to use the Monroe tests. The tests may, however, be criticized on the basis of the timing of the rate of dictation. Although the rate of dictation is slightly slower than the average writing rate of children as determined by tests of 6,000 children in each of the school grades, it is too fast for many children. Ayres⁸ has shown the great overlapping in speed of handwriting within any school grade. For example, in the fifth grade some children write twenty times faster than other children in the same grade. Approximately 31 per cent of the eighth grade children write no more rapidly than the average fifth grade child. Monroe believes that this is not a serious fault of the scale, since the words from the Ayres list are placed in the earlier parts of the sentences. This is only partially true. It is certain that many slow writers will make low scores on these tests not because of poor spelling ability, but because of slow writing. Whenever possible any factor to be measured should be isolated from other complicating factors. In these tests the measurement of spelling ability is complicated by the introduction of the factor of speed of handwriting.

SAMPLE SENTENCES FROM MONROE'S TIMED SPELLING TEST
FOR THE FIFTH GRADE

Seconds

- 60 The *president* gave *important information* to the men.
- 48 The *women* were present at the time.
- 19 The *entire* region was *burned over*.
- 49 The *gentlemen* declare the *result* was *printed*.
- 30 *Suppose* a *special attempt* is made.

⁸ See, "Ayres Handwriting Scale," *Russell Sage Foundation*, New York City.

IOWA SPELLING SCALE

Description and derivation.—This test was devised by Ernest J. Ashbaugh especially for the measurement of the spelling ability of elementary schools of Iowa.⁹ The scale includes 2,977 words taken from the written correspondence of Iowa people. The difficulty of the different words was determined by a total of about 4,662,200 spellings by children from each of the school grades. The words are arranged into twenty-five groups, or "steps," each group varying from the one just above or below by approximately equal differences in difficulty. The scale is divided into three parts: part one for use in the second, third and fourth grades; part two for use in the fourth, fifth and sixth grades; and part three for the sixth, seventh and eighth grades.

Methods of using and scoring the scale.—The author suggests that the teacher may use the scale in any one of three ways: (1) It may be used as a minimal list of words which the children of the elementary grades should be taught. (2) It may be used as an instrument for measuring the comparative skill with which children can spell a certain list of words as compared with the average for the State of Iowa. Norms for each grade are given for each step or group of words. Used for this purpose it is a real spelling scale. (3) It may be used as a measure in teaching. By comparing scores from time to time a teacher can measure her success in teaching spelling.

Function of the scale.—This spelling scale presents a very practical means of providing the class-room

⁹ University of Iowa Extension Bulletin, Nos. 53, 54, 55, Iowa City, Iowa.

teacher with both a measuring scale and subject-matter in spelling. The number of words is large enough to include practically all the common words in a child's vocabulary. The method for using the scale is the same as that ordinarily used by the teacher in written spelling. The teacher needs only to compare the scores of a pupil or a class with the norm to determine whether the pupil or class is above or below the given standard in spelling ability.

The Iowa Spelling Scales are very similar to the Ayres Scale. About three times as many words are included in the former, but both have much in common with respect to the selection and placement of words.

MATERIAL OF ENGLISH SPELLING—JONES

Description and derivation.—W. F. Jones made a study similar to that of Ayres except that the sources of his material were compositions written in school. He listed the words used by 1,050 children, approximately 150 from each grade, from four schools in widely separated parts of the United States. In all about 15,000,000 words were listed, but this list represents only 4,532 different words. These words are arranged in the Jones list by grades by including each word in the lowest grade in which at least 2 per cent of the pupils used it.

From this list Jones selected the 100 words most often misspelled and arranged them into a list called the "One Hundred Spelling Demons of the English Language." Children should receive special drill in the correct spelling of this list of common words.

which	can't	guess	they
their	cure	says	half
there	loose	having	break
separate	lose	just	buy
don't	Wednesday	doctor	again
meant	country	whether	very
business	February	believe	none
many	know	knew	week
friend	could	laid	often
some	seems	tear	whole
been	Tuesday	choose	won't
since	wear	tired	cough
used	answer	grammar	piece
always	two	minute	raise
where	too	any	ache
women	ready	much	read
done	forty	beginning	said
hear	hour	blue	hoarse
here	trouble	though	shoes
write	among	coming	to-night
writing	busy	early	wrote
heard	built	instead	enough
does	color	easy	truly
once	making	through	sugar
would	dear	every	straight

TEACHERS WORD BOOK—THORNDIKE

Description and derivation.—*The Teachers Word Book* was compiled by Edward L. Thorndike to represent a more complete and satisfactory list of the most commonly used words in the English language. It consists of ¹⁰ “an alphabetical list of 10,000 words

¹⁰ See the introduction to *The Teachers Word Book*, E. L. Thorndike, (Bureau of Publications, Teachers College, Columbia University, New York City).

which are found to occur most often in a count of (1) about 525,000 words taken from the literature for children, (2) about 3,000,000 words from the Bible and English classics, (3) about 300,000 words from elementary school textbooks, (4) about 50,000 words from books about cooking, sewing, farming, the trades, and the like, (5) about 90,000 words from daily newspapers, and (6) about 500,000 words from correspondence." In all, forty different sources were used.

The words are listed alphabetically and their relative frequency indicated by numbers at the side. The 1,000 words used most frequently have a credit number of forty-nine or more uses.

The values of such a list of words as set forth by Thorndike are: (1) To inform the teacher that words in the reading lesson should receive the most attention. Many words are found in the readers that are not used frequently enough to warrant special study. The Word Book gives the teacher a method for selecting the words in the lesson for careful study. (2) It may be used by the less experienced teacher to provide her with that knowledge, both of the importance of words and of their difficulty, which the expert teacher has acquired by years of experience with pupils and books. (3) The Word Book may be used as a convenient place to record any useful facts about the words contained therein. (4) It may be made the basis for the construction of spelling lists of the most common English words. In fact it is the most carefully selected and complete list of its kind in the English language.

MORRISON-McCALL SPELLING SCALE

Description and derivation.—This scale is composed of eight lists of fifty words each, all lists being of equal difficulty. The lists of words are based upon the Ayres Scale, the Buckingham extension of the Ayres, and the Thorndike Word Book. The words were selected from the first two named scales in such a manner as to make all the lists of equal difficulty; and, in addition, the words had to be among the 5,000 most commonly employed, as indicated in the Thorndike Word Book.

Method of giving and scoring the scale.—Any one of the eight lists may be used for testing purposes, inasmuch as they are all equally difficult. The word to be spelled is first pronounced, then used in an illustrative sentence which is prescribed in the booklet, and finally pronounced again. The time required will vary with the grade, inasmuch as the words and sentences are to be read at a rate which seems best suited to the class, thereby eliminating the speed factor.

Each word is either right or wrong, the standard being absolute accuracy. The score is the number of words spelled correctly. The norms were derived from testing 57,337 pupils in rural and village schools, using about 8,000 pupils in grades 2 to 8 and about 1,000 in grade 9.

GRADE NORMS IN TERMS OF AVERAGE NUMBER OF WORDS
SPELLED CORRECTLY

(mid-year averages)

Grade	2	3	4	5	6	7	8	9
Average number of words .	11	18	24	30	35	39	42	44

Function of the scales.—The scales are intended to indicate a pupil's spelling ability relative to other pupils of the same grade and age, and to indicate the grade and age for which the pupil's spelling ability is "normal." In the same way a class average may be compared. For these purposes, the authors include in their description not only the above stated norms, but also tables for finding the so-called T-scores,¹¹ and tables for finding age and grade status of an individual, or the status of a group.

Summary.—It is clear from the descriptions of the tests herein considered that they have been constructed on a principle of utility, for they are made up of words which occur in nearly all of the more common situations where spelling is essential. The lists, of course, are samplings; but these, when properly selected and scaled, may be highly reliable. The use of standardized tests in spelling makes it possible to determine, with a much closer approach to accuracy, the relative merits of different methods of teaching the subject. Furthermore, the nature of the words included in the spelling tests gives the added assurance that the time spent on spelling will be of value, inasmuch as the words will form part of the pupils' active vocabularies.

MATERIALS NEEDED

Ayres Spelling Scale for grades 2 to 8. Only one copy of scale needed. Price 5 cents. (Russell Sage Foundation, New York City.)

¹¹ A standard scoring device which takes the score of the child 12 years and 6 months of age as a single norm for uniform comparisons. See "Uniform Method of Scale Construction," (*Teachers College Record*, January, 1921).

Buckingham Extension of the Ayres Spelling Scale for grades 2 to 8. Only one copy needed. Price 14 cents. (The Public School Publishing Company, Bloomington, Ill.), *Spelling Ability, Its Measurement and Distribution*, B. R. Buckingham (Bureau of Publications, Teachers College, Columbia University, New York City).

Iowa Spelling Scale (Ashbaugh) for grades 2 to 8. Only one each of the three scales needed. (See University of Iowa Extension Bulletin, Nos. 53, 54 and 55.)

Jones Spelling Scale for grades 2 to 8. Only one copy necessary. See *Concrete Investigation of the Material of English Spelling* by N. F. Jones (University of South Dakota, Vermillion, S. D., 1913).

Monroe Timed Spelling Tests for grades 3 to 8. Test No. 1 for grades 3 to 4, test No. 2 for grades 5 to 6, test No. 3 for grades 7 to 8 and high school. Only one copy of the test needed for use in the grades indicated. Price 4 cents per test or single set 12 cents. (The Public School Publishing Company, Bloomington, Ill.)

Morrison-McCall Spelling Scale, for grades 2 to 8. One copy for each examiner needed; no pupil material necessary. Price 30 cents per copy. (World Book Company, Yonkers-on-Hudson, N. Y.)

The Teachers Word Book (Thorndike) (Bureau of Publications, Teachers College, Columbia University, 1921).

SUPPLEMENTARY LIST OF TESTS

Courtis Standard Research Tests in Spelling.¹² Tests for each half grade from II-B to VIII-A. Two forms of each, one for beginning and one for end of semester. Based on Ayres' list. National Spelling Scales.¹³ Four forms for elementary schools and four for junior high schools. Based on the Buckingham Extension, the Seven S Spelling Scales, the Iowa Scales, and the Thorndike Word Book.

¹² Courtis Standard Tests, 1807 East Grand Boulevard, Detroit, Mich.

¹³ National Publishing Society, Mountain Lake Park, Maryland.

Sixteen Spelling Scales.¹⁴ Standardized in sentences for secondary schools

Tidyman Standard Spelling Tests.¹⁵ For use with Supervised Study Speller, to measure initial ability, progress, and to make comparisons. Grades 2 to 8.

Van Wagenen Spelling Scales.¹⁶ For grades 3 to 8; all in one booklet.

SELECTED REFERENCES

Anderson, W. N., "Determination of a Spelling Vocabulary based upon Written Correspondence" (*University of Iowa Studies in Education*, Vol. 2, No. 1).

Ayres, L. P., "A Measuring Scale for Ability in Spelling" (New York, *Russell Sage Foundation*).

Buckingham, B. R., *Spelling Ability: Its Measurement and Distribution* (Bureau of Publications, Teachers College, Columbia University, 1918).

Cornman, O. P., *Spelling in the Elementary Schools* (Boston, Ginn and Company, 1902).

Hollingsworth, L. S., *The Psychology of Special Disability in Spelling* (Bureau of Publications, Teachers College, Columbia University, 1918).

Horn, E., "A Source of Confusion in Spelling" (*Journal of Educational Research*, Vol. 19, No. 1, pp. 47-55).

Horn, E., "Ten Thousand Words Most Commonly used in Writing" (*Monograph, State University of Iowa*. Cf. *Third and Fourth Yearbooks of the Department of Superintendence*).

Hudelson, Earle, and others, "Sixteen Spelling Scales Standardized for Use in Secondary Schools" (*Teachers College Record*, Vol. 21, p. 337 ff., 1920).

Reed, H. B., *Psychology of Elementary School Subjects* (Boston, Ginn and Company, 1927).

¹⁴ Bureau of Publications, Teachers College, Columbia University.

¹⁵ State Teachers College, Emporia, Kansas.

¹⁶ The Educational Test Bureau, University of Minnesota.

Rice, J. M., "The Futility of the Spelling Grind" (New York, *The Forum*, 1897, pp. 163 and 409).

Shambaugh, C. G., and Shambaugh, O. L., "A Core Vocabulary for Elementary School Pupils" (*Journal of Educational Research*, Vol. 19, No. 1, pp. 39-46).

Starch, D., *Educational Psychology*. (Revised Edition) Chapter XIX (New York, The Macmillan Company, 1927).

Suzzallo, H., "The Teaching of Spelling" (*Teachers College Record*, Vol. 12, No. 5).

Tidyman, W. F., *The Teaching of Spelling* (Yonkers-on-Hudson, World Book Company, 1926).

Wallin, J. E. W., *Spelling Efficiency in Relation to Age, Grade, Sex, and the Question of Transfer* (Baltimore, Warwick & York, 1911).

CHAPTER VI

HANDWRITING

Problems in the measurement of handwriting.—Handwriting, though largely a drill subject, is a very complex process, involving visual-muscular coördination. Psychologically, the learning of writing is the development of muscular habits which will result in handwriting having legibility, speed, and perhaps esthetic quality. In other words, there are two principal factors to be considered in the measurement of handwriting: legibility (quality) and speed (quantity). The latter of these may be readily ascertained by noting the number of letters written in a given unit of time. But the great difficulty has been the measurement of quality and the unit to be employed. Handwriting, more than many other school subjects, demonstrates the inadequacies of subjective judgment, for teachers differ markedly with respect to what constitutes good handwriting. And, as is to be expected under such conditions, the same specimen will receive a wide range of gradings from different teachers, and the relative merits of several specimens will not be agreed upon. Furthermore, it has been demonstrated that the same teacher will not assign the same marks to the specimens on reëxamining them.

Quality in handwriting.—In the construction of a handwriting scale, the first problem is the determination of what constitutes successive stages of quality. That is to say, what sort of specimen should be rated

30, or 60, or 70? As will appear later, this question has been answered to a high degree of satisfaction by current measuring scales which consist of graded samples separated by approximately equal differences in merit or quality. Measurement in handwriting consists in placing the sample alongside the scale and noting which step, or level, it resembles most closely in quality. The scale may consist of a number of steps having given numerical values, such as the eight steps in the Ayres Scale designated as 20, 30, up to 90. By use of such a measure the teacher may transfer uncertain and variable notions of quality into objective reality.

Ordinarily, the examiner fails to distinguish between the various levels of quality of handwriting to be evaluated. This was shown by Ayres¹ in a study of ratings given applicants on a Civil Service examination. When the papers were re-scored by the use of a scale, it was found that while the examiners had ranged the grades from 60 to 95, the papers varied from 20 to 90, on the basis of the scale standards. The following table gives a comparison of the two sets of grades:

<i>Quality as rated by Civil Service Examiners</i>	<i>Quality as measured on the Ayres Scale— based on legibility</i>
60	20
65	30
70	40
75	50
80	60
85	70
90	80
95	90

¹ Ayres, L. P., "A Scale for Measuring the Handwriting of Adults," (Russell Sage Foundation, New York City).

Speed in handwriting.—Another matter arising in connection with the question of handwriting is the problem of rate, or speed. The problems of speed and quality are interdependent and should be measured simultaneously in the same test, inasmuch as these two aspects of writing are functionally related. It is desirable, therefore, that a test of writing be made by requiring the pupils to write a sentence or a short passage repeatedly as many times as possible within a short period of time—say, two or three minutes. Quality may then be measured by one of the scales, and speed by the number of letters written per minute. Undue speed should not be sought at the sacrifice of quality; nor, on the other hand, should quality alone be stressed to such an extent that reasonable speed is lost. Tests and comparisons with standards will reveal to the teacher whether one or the other of these aspects is being sacrificed. Experiment has shown that it is best, in teaching handwriting, to stress both quality and speed, for it has been demonstrated that only to a very slight degree is the good writer extremely slow and the rapid writer extremely poor.²

The several handwriting scales are supplied with grade norms, so that by comparing her class average on an individual's scores with the norms the teacher may learn whether her pupils are at standard, or whether they deviate above or below. In case they are below standard, by the use of a diagnostic scale she may discover in what specific respects they are poor and concentrate drill on these. In case the pupils are

² Starch, D., "The Measurement of Efficiency in Handwriting" (*Journal of Educational Psychology*, Vol. 6, pp. 106-114).

above normal the teacher may well spend less time on handwriting and devote more time to other subjects.

The question naturally arises, "What level of proficiency should be attained in handwriting?" Tests conducted in a number of schools have shown that the average attainment in writing at the end of the eighth grade is as good as quality 60 on the Ayres Scale, or quality 11 on the Thorndike Scale, while the average speed is about 83 letters per minute. In order to answer the question of whether this level of proficiency was adequate for most purposes, several investigations have been made which demonstrated that writing which is the equivalent of Ayres quality 60 is sufficiently good for nearly all purposes.³ There are, however, a small number of occupations for which the quality should be of a higher level, but probably not in excess of 70. Among these are elementary school teachers, clerks, bookkeepers. The evidence further points to the conclusion that at the end of the sixth grade a child should achieve a rate of 70 letters per minute and a minimum quality of 50 on the Ayres Scale. According to one investigator, additional drill in writing should be given those children who pass into the seventh grade without having attained these minimum scores.⁴

One caution should be given to the teacher in her first use of a handwriting scale. She may find almost as great variation in two markings of the same papers

³ Freeman, F. N., "Handwriting" (*Fourteenth Yearbook of the National Society for the Study of Education*, 1915).

Koos, L. V., "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools" (*Elementary School Journal*, Vol. 18, 1925, pp. 423-446).

⁴ Freeman, F. N., "Handwriting" (*Third Yearbook, Department of Superintendence, N. E. A.*, 1925, pp. 205-216).

by herself, or between her markings and the markings of the same papers by another teacher, with the use of the scale as without. But a short period of practice in the use of the scale, especially by a group of teachers discussing the points of variability, will soon produce remarkably little variability in the assignment of grades to specimens of handwriting by a teacher, and little variation between the grades of different teachers. C. T. Gray⁵ reports an experiment in which three students who had had no experience in teaching or in the use of scales were given practice in grading samples of handwriting by means of the Ayres Scale. Twenty-five samples were graded each week by each person. After the grading, conferences were held to compare grades and discuss difficulties in grading. The average variation between the highest and the lowest grades for the twenty-five samples for the first week was 20.4. This was considerable variation, but probably not more than would have been found in grading without a scale. By the fifth week the variation was reduced to 12.7, and for the fifteenth week it was only 3.6. It seems from this experiment that a person without experience, by grading three or four hundred specimens of handwriting, could reduce the variability to an almost negligible factor. No doubt the average teacher accustomed to grading could do as well in a shorter time. If this be true, the results would highly justify the small amount of effort required.

⁵ Gray, C. T., "The Training of Judgment in the Use of the Ayres Scale for Handwriting" (*Journal of Educational Psychology*, Vol. 6, 1915, pp. 85-95).

Another study by one of the authors⁶ further confirms Gray's results. By way of contrast, however, Gilliland found that when there is no objective standard with which to compare the samples, practice will not improve ability in grading. With materials selected from Thorndike's samples, a group of students deviated from the true scores on an average of 10.9, 8.9, and 9.8 per cent in three successive evaluations of the samples. This grading was conducted by the usual subjective methods. Even on the first application of a standard scale, these same judges reduced their average error to 5.6 per cent. Through practice in evaluating with a scale other samples of handwriting with known scores and then checking the results with the known scores, these judges were able to reduce their errors in grading the original samples to 4.0 and 2.6 per cent respectively in the next two attempts. This demonstrates the advantage of using objective standard tests in grading handwriting.

With this general discussion of handwriting scales let us pass to a study of some of the more important scales.

FREEMAN ANALYTICAL HANDWRITING SCALE

Description of the scale.—The scale consists of five separate parts, one for measuring each of the five elements of handwriting. The elements are (1) uniformity of slant, (2) uniformity of alignment, (3) quality of

⁶ Gilliland, A. R., "The Effect of Practice with and without Knowledge of Results in Grading Handwriting" (*Journal of Educational Psychology*, Vol. 16, pp. 539-547, 1925).

line, (4) letter formation, and (5) spacing. The scale consists of three samples representing different degrees of quality in each of the five elements. The poorest sample in each case is given a grade of 1, the medium sample a grade of 3, and the best sample a grade of 5. The intermediate scores, 2 and 4, may be used.

Method of using the scale.—Any sample of handwriting to be graded is scored on each of the five elements of the scale separately and independently of the others. The sample is first given a grade between 1 and 5 on uniformity of slant by comparing it with the three samples for this part of the scale. In the scoring of this aspect of handwriting, all factors other than the uniformity of slant are to be entirely disregarded by the grader. After this scoring is completed, the paper may be graded in each of the other four elements of the scale in turn. The final score for any paper is the sum of the individual scores on each of the five elements. When the specimens of a group of pupils are being rated, it is advisable to grade *all* the papers on one of the elements at a time. By so doing greater uniformity of rating is assured.

The speed of writing is measured as well as quality. Freeman gives the following norms in speed and quality for grades 2 to 8.

NORMS FOR FREEMAN HANDWRITING SCALES

Grade	2	3	4	5	6	7	8
Quality Score ⁷	17.9	18.4	19.0	20.0	20.8	22.0	23.0
Speed (letters written per minute) ⁸	36	48	56	65	72	80	90

⁷ *The Teaching of Handwriting*, (Boston, Houghton Mifflin Company, Chap. V).

⁸ *Fourteenth Yearbook*, National Society for Study of Education.

Function of the scale.—In many respects the Free-man Handwriting Scale is an ideal measuring scale. The characteristics of good or poor handwriting have been carefully analyzed. Each character is independently measurable by reference to given standards. This makes the scale objective. Definite norms have been determined for both speed and quality; but more important still for the class-room teacher is the fact that the scale is diagnostic. In other words, the teacher by using this scale may find out not only the bare fact that a pupil or grade is below or above standard, but also in what elements or qualities the pupil or pupils are superior and in which they fail. It is very important for a teacher to know, for example, that a child's handwriting is good in quality of letters but poor in alignment and letter spacing. Knowing these facts the teacher may undertake remedial work; she may point out to the pupil his own defects, and the pupil may in turn, by noting his defects and by comparisons with standards, progress far in his own improvement. Furthermore, the scale is diagnostic in the sense that the factors of speed and quality are separated, so that the examiner may detect whether one is being sacrificed to the other. To the teacher, of course, falls the task of supplying such remedial work as will overcome a pupil's deficiencies.⁹ Important as this task is, it is beyond the province of this text more than to point out its significance and refer the reader to treatises on remedial teaching for further discussion.

⁹ A recent book on the subject is *How to Teach Handwriting*, by Free-man, F. N. and Dougherty, M. L., (Boston, Houghton Mifflin Company, 1923).

Uniformity of Slant

A quick brown fox
quite brown fox jumps

Uniformity of Alignment

A quick brown fox jumps over
A quick brown fox

Quality of Line

A quick brown fox jumps over
A quick brown

Letter Formation

A quick brown fox
A quick brown fox jumps

Spacing

A quick brown fox jumps over
A quick brown fox jumps over the

1

Uniformity of Slant

Some books are to be slanted, others to be
Some books are to be slanted, others

Uniformity of Alignment

A quick brown fox
parts, others to be read but no

Quality of Line

A quick brown fox jumps
Some books are to be

Letter Formation

A quick brown fox
A quick brown fox jumps over

Spacing

A quick brown fox jumps over
A quick brown fox jumps over the

The Freeman scale has not been used as widely as some of the other handwriting scales. This is probably due to the fact that it has been employed more often as a measure of general merit of handwriting rather than as an instrument to diagnose faults in handwriting for which it is particularly suited. There are other instruments which are perhaps better adapted to the problem of rating the merit of handwriting.

AYRES MEASURING SCALE FOR HANDWRITING

Description and derivation of the scale.—The latest form of the Ayres Scale, called the Gettysburg Edition, consists of a series of eight samples of handwriting varying by increments of ten from the poorest sample with a value of 20 up to the best with a value of 90. Each sample consists of the first few lines of Lincoln's Gettysburg Address.

The Gettysburg Edition of the Ayres Scale is the outgrowth of an earlier scale which is known as the Three Slant Edition. It was devised in 1912 as the result of the study of about 1,500 samples of children's handwriting taken from representative schools in 38 states. These samples were ranked by ten investigators on the basis of the time required to read the selection. That is, the quality of each sample of handwriting was determined by the degree of legibility as shown by the rate of reading by these ten judges.

Three sets of samples were then selected and assigned equivalent values of 20, 30, 40, 50, 60, 70, 80, and 90. One set was written in ordinary slant, another set

in vertical writing, and the other in backhand. The successive steps of the samples represent uniform increments of legibility in writing; thus the superiority of a sample rated 50 over one rated 40 is the equivalent of the superiority of an 80 sample over one rated 70.

The Gettysburg Edition does not supersede the Three Slant Edition. It is similar in nature and construction to the latter, except that there is only one sample of each quality, written with a medium slant. The purpose of the Gettysburg Edition is to increase the reliability of the measurements through standardization of the methods of securing and scoring the samples. The authors believe that the new scale will reduce variability in results.

Method of using the scale.—The children to be graded in handwriting are drilled on the first three sentences of the Gettysburg Address until they are familiar with them. They are then provided with pen, ink, and ruled paper. At a given signal they begin to write and continue for two minutes. The papers are then collected for scoring.

The samples are scored on quality by passing each paper in turn along a copy of the scale until a point is found where the quality of the sample and the quality of the scale are matched most closely. Differences in style are disregarded. The accompanying score on the scale is given as the grade for the sample of handwriting. The rate of writing is determined by finding the average number of letters written per minute during the two-minute writing period.

20	30
<p>Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war testing whether that nation or any nation so conceived and so dedicated, can</p>	<p>Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war testing whether that nation or any nation so conceived and so dedicated, can</p>

This scale for measuring the quality of handwriting is a revised edition of a scale first published in 1912 and subsequently reprinted 12 times with several minor revisions and with a total of 62,000 copies. The purpose of the changes introduced in the present edition is to increase the reliability of measurements of handwriting through standardizing methods of securing and scoring samples, and through making numerous improvements in the scale itself designed to reduce variability in the results secured through its use. The present scale may be referred to as the "Gettysburg Edition" in order to distinguish it from other editions. The original or "Three Slant Edition" and the scale for adult handwriting are not superseded by the present scale. Copies of any of the three scales may be secured for five cents each, postpaid.

To secure samples of handwriting the teacher should write on the board the first three sentences of Lincoln's Gettysburg Address and have the pupils read and copy until familiar with it. They should then copy it, beginning at a given signal and writing for precisely two minutes. They should write in ink on ruled paper. The copy with the count of the letters is as follows.

Year 4 score 9 and 12 seven 17 years 22 ago 25 our 28 fathers 35 brought 43 forth 47 open 31 that 33 contained 64 x 45 new 68 nation 74 deceived 83 in 85 liberty 92 and 85 dedicated 104 to 106 the 109 proportion 120 that 124 all 127 men 130 are 123 created 140 equal 145. Now 148 we 150 are 153 engaged 160 in 162 a 163 165 168 170 172 war 176 testing 183 whether 190 that 194 nation 200 or 202 copy 205 nation 211 no 221 concealed 222 and 25 so 227 dedicated 236 can 238 long 245 endorse 249. We 251 are 254 met 237 on 259 a 260 great 265 battlefiled 276 of 278 that 282 war 285.

SECTIONS OF THE LOWER END OF THE GETTYSBURG EDITION OF THE AYRES HANDWRITING SCALE

Norms for both speed and quality are as follows:

AYRES HANDWRITING NORMS

Grade	2	3	4	5	6	7	8
No. of letters written per min.	31	44	55	64	71	76	79
Quality score	38	42	46	50	54	58	62

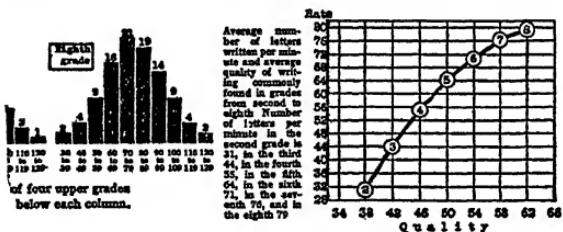
From the norms it may be seen, for example, that a fourth grade class should write at an average rate of about 55 letters per minute, with an average quality score of between 40 and 50. In the same manner the achievement of an individual may be compared.

80

Four score and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing

90

Four score and seven years ago our fathers brought forth upon the continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing



Division of Education
Russell Sage Foundation
130 East 22nd Street, New York City
LEONARD E. AYRES, Director

SECTIONS OF THE UPPER END OF THE GETTYSBURG EDITION OF THE AYRES HANDWRITING SCALE

Function of the scale.—The use of this scale makes the grading of handwriting highly objective. As has already been indicated, a relatively small amount of practice makes the average class-room teacher very proficient in the use of the scale. It is simple to administer, and the norms are very reliable. Some teachers may object to making legibility the basis for scoring handwriting. If so, the Thorndike Scale should be used. But practically, at least, no other criteria are so important or satisfactory as legibility as a basis of merit.

THORNDIKE SCALE FOR THE HANDWRITING OF CHILDREN IN GRADES FIVE TO EIGHT

Description and derivation of the scale.—This scale, constructed by Professor E. L. Thorndike, consists of 29 samples of handwriting ranging in quality from 4 to 18. There is a sample for each step on the scale and in many cases more than one sample for a step, representing different styles of handwriting. Steps in the scale represent differences in general merit of handwriting as described later.

The material for qualities 5 to 17 of the scale was taken from actual samples of the handwriting of children. The sample for quality 4 was artificially constructed and that for quality 18 was taken from a copy book. There were about 1,000 samples in all. These samples were rated by from 23 to 55 judges on the basis of general merit ¹⁰ of handwriting. Slant, style, or other special factors were not taken into consideration in the rating.

While there is no sample that was ranked as zero, this point was defined "roughly as handwriting recognizable as such but of absolutely no merit as handwriting." This zero point is of theoretical interest, at least, for, as Thorndike has pointed out, three things are necessary for any kind of measurement: (1) a zero or beginning point, (2) an ending point, and (3) a unit of measure. Thorndike uses a rather complicated statistical method of arriving at his unit of measure, but he so constructs it that the difference between sample 4 and sample 5, for instance, is the same as the differ-

¹⁰ Thorndike, E. L., "Handwriting" (*Teachers College Record*, Vol. 2, No. 2, March, 1910).

ence between sample 17 and sample 18; so that sample 8 is just twice as good as sample 4. "The unit of the scale equals one-tenth of the difference between the best and the worst of the formal writing of 1,000 children in grades 5 to 8."

Method of using the scale.—The scale is to be used by comparing the specimens of handwriting to be measured with the samples in the scale and assigning a score on the basis of this comparison from the scores given on the scale. Fractional scoring between the units of the scale is allowed. If it is desired to grade on the basis of 100, the Thorndike score may be transposed to this basis by multiplying it by 5.5. For example, a score of 12 on the Thorndike Scale is equivalent to a score of 66 on the scale of 100.

We have already considered the problem of the training of the teacher in the use of a handwriting scale. Thorndike has provided a valuable means whereby the teacher may improve her scoring in handwriting by supplying fifty samples of handwriting with the true values of the specimens graded by the Thorndike Scale.¹¹ The teacher may practice grading these specimens, and by comparing her scores with the true scores, she may determine the direction and amount of her errors in scoring. Thorndike says that an average competent teacher who is without training in the use of the scale will make an error of .9 (4.95 on the scale of 100) of a step in judging a sample. Practice on the fifty specimens with knowledge of the results should lower this error.

¹¹ "Teachers' Estimates of the Quality of Specimens of Handwriting," by E. L. Thorndike, (*Teachers College Record*, Vol. XV, No. 5, November, 1914).

Quality 7. Sample 126

card, John vanished behind the bushes and the carriage moved

Quality 6. Sample 12

gathering about them melted away in an instant leaving only a poor old lady

Quality 5. Sample 6

bushes and the carriage swerved along down the driveway. ~~see~~ another

Quality 4. Sample 121

sat on the curb was my driver and

Function of the scale.—The Thorndike Scale is well suited for the purpose for which it is intended, that is, as an aid to the teacher in scoring handwriting on the basis of general merit. The scale is not diagnostic as it does not indicate the elements of merit or demerit. Norms have not been derived for the different school grades.¹² Speed of writing is not taken into account in this scale. Its purpose is to present an objective standard as the basis for grading handwriting.

GRAY STANDARD SCORE CARD FOR MEASURING HANDWRITING

This is not a test or scale but, as its name indicates, a card for recording certain qualities of handwriting. It is constructed on the same general principle as score cards used in judging stock and fruit and is to be used in the same way and for the same general purpose. It is designed to point out the good and the poor qualities in handwriting. There are nine separate qualities listed on the card. The highest possible score varies from 3 for heaviness of line to 26 for general form; 100 points represent a perfect score.

Like the Freeman Scale this score card is valuable for pointing out the different qualities that go to make up handwriting and the relative importance of each. Too often the pupil and even the teacher think only in terms of general merit without any very definite idea

¹² Starch gives norms for the Thorndike Scale based on a study of 6,000 pupils in 28 schools. Daniel Starch, *Educational Psychology*, p. 352.

Grade	1	2	3	4	5	6	7	8
Speed	20	31	38	47	57	65	75	83
Quality	6.5	7.5	8.2	8.7	9.3	9.8	10.4	10.9

of what constitutes merit. The teacher may make use of the score card for grading her pupils in handwriting and in this way it really becomes a diagnostic scale.

COURTIS STANDARD PRACTICE TESTS IN HANDWRITING

Description and derivation of the tests.—These are not tests in the ordinary sense of the word but are a combination of tests and practice exercises in handwriting. The tests and the method of using them are described in a Student's Daily Lesson Book and a Teacher's Manual. The Daily Lesson Book provides each child with copies for practice lessons and lesson helps to go with each copy. The Teacher's Manual describes the tests, how the pupils are to use them, how the teacher is to help the children in their use, how to measure progress through the use of the tests, and what use the teacher should make of the results.

These tests were devised by S. A. Courtis and Lena A. Shaw of Detroit as a result of three years' experimentation in handwriting. A series of twenty lessons chosen from a textbook in business writing was arranged so that the easiest forms came first and the more complex forms followed. The lessons progressed from simple words to more difficult words and phrases, then whole sentences and paragraphs. These lessons were supplemented later by others and the material rearranged in a more logical form.

In a later arrangement, the material for the first lessons was determined by a study of the relative frequency with which various letters of the alphabet occur in everyday usage, except that this principle was

somewhat modified because of differences in difficulty of the formation of some letters. Standard rates and qualities for the Courtis Tests were derived from a study of 1,000 samples of handwriting. The quality of these specimens was measured by and the norms given in terms of the Ayres Scale.

Method of using the tests.—“On the first day a research test is given in order to find out what children need drill, what kind of drill and how much, and whether there are children in the class who would not profit by working on the practice tests. Those who fail to pass the research test begin on Lesson One, and each child must continue on that lesson until he has written it fast enough and of a quality up to standard for his grade before he can pass to the next lesson. The result is that children work on those lessons on which they need to work. The child who can progress at the correct rate does not need the help of his teacher. The child who goes too slowly receives the individual attention of the instructor. Provision is made for the teacher to discover his slow progress, or lack of progress, and the weakness in his writing. She then helps to remedy them.

“The child scores his paper and enters his own record and graph after every day’s work. As a result of these two operations on his part he learns to judge for himself why his work is not so good as it should be and how much more rapidly he should do his work in order to make it equal, or surpass, that which is set as the standard for his grade.

“The children who do not need drill from the beginning of the work or those who finish the series of les-

sons in a very short time may be excused from practice work in handwriting for that grade, or they may be given the standard for the next grade, whichever method seems advisable in the particular school system in which the child is working.”¹⁸

In general each day’s work is to be divided into three parts: (1) special practice, 5 minutes, (2) testing, 5 minutes, (3) scoring and recording, 5 minutes. The purpose of the practice tests is “to teach the children to teach themselves to write well.” As already stated the pupils are to grade their own specimens of handwriting. Quality is scored by the use of the Ayres Scale. Rate is the number of letters written in three minutes. The following norms are given:

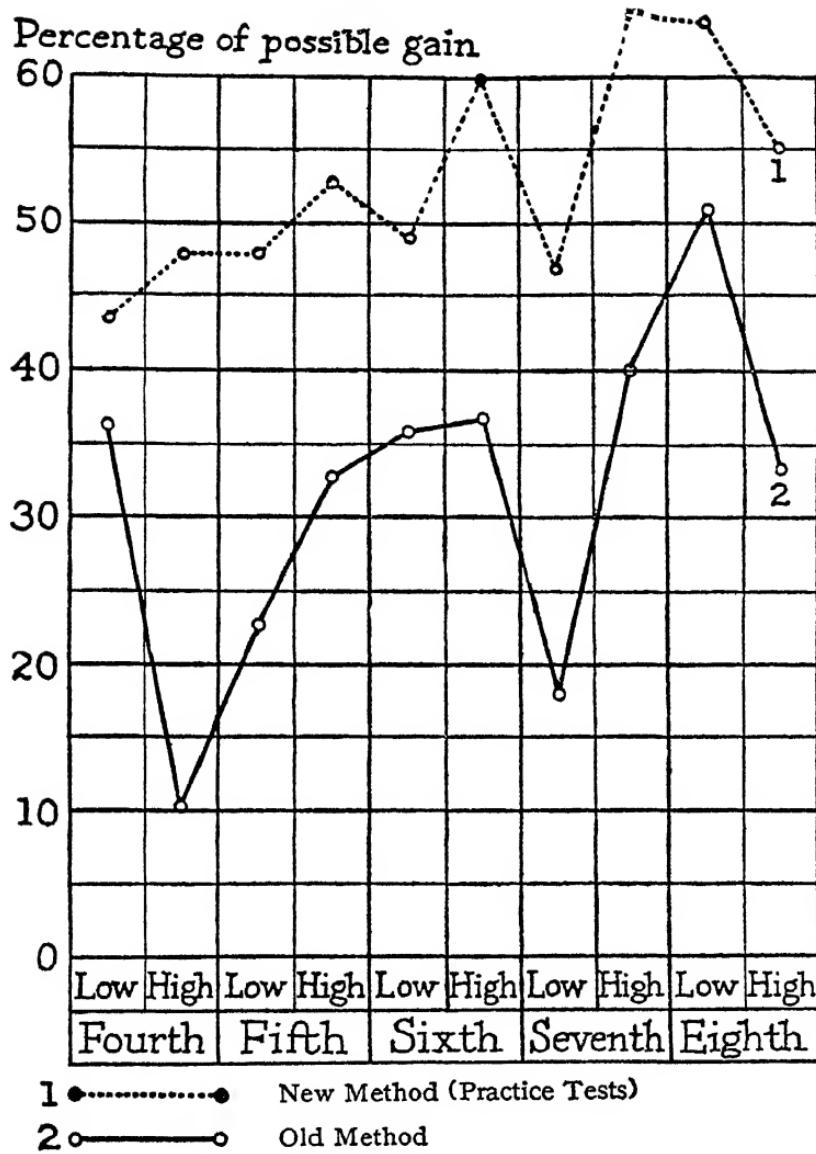
COURTIS HANDWRITING NORMS

<i>Grade</i>	3	4	5	6	7	8
<i>Standard</i>	low, high					
<i>Rate</i> ...	40	46	52	58	62	66
<i>Standard</i>	69	72	73	78	80	82
<i>Quality</i> ...	45	50	55	60	65	70

Thus a pupil in the high sixth grade should write 72 or more letters in three minutes and the quality should be at least 60 on the Ayres Scale.

In order to test the value of the use of the tests, Courtis arranged two groups of pupils, one of which used the practice tests and the other received the ordinary instruction in handwriting. The two groups were so arranged that their handwriting was about equal at the beginning of the experiment. The differences between the two groups at the end of the school year are shown in the accompanying graph.

¹⁸ Courtis, S. A., Bulletin No. 1, Courtis Standard Practice Tests in Handwriting, p. 9.



FROM COURTIS BULLETIN, No. 1, P. 9

Fig. 1

"In every class, except one, the pupils who used the Courtis Standard Practice Tests in Handwriting made a greater percentage of gain than those who studied penmanship without them, the least gain being a little less than 20% better, and the greatest gain being as much as 380% better." The beginning teacher, or the grade teacher without a supervisor, can make especially good use of such a carefully constructed teaching device as these practice tests in handwriting provide.

MATERIALS NEEDED

Ayres Handwriting Scale, Gettysburg Edition for grades 2 to 8. Price 10 cents. (Russell Sage Foundation, New York City.) Courtis Standard Practice Tests in Handwriting. Specimen Set including Student's Daily Lesson Book, Student's Daily Record Card, Teacher's Manual, Ayres Handwriting Scale, Gettysburg Edition (2 copies), all for grades 3 to 8. Price 45 cents. (Yonkers-on-Hudson, N. Y., World Book Company).

Freeman Chart for Diagnosing Faults in Handwriting for grades 2 to 8. Price 30 cents. (Boston, Houghton Mifflin Company.)

Gray Standard Score Card for Measuring Handwriting; for grades 2 to 8. Sample 15 cents. (The Public School Publishing Company, Bloomington, Illinois.)

Thorndike Handwriting Scale, for grades 5 to 8. (Bureau of Publications, Teachers College, Columbia University.)

SUPPLEMENTARY LIST OF SCALES

Graves Diagnostic Chart for Handwriting.¹⁴

Graves Measuring Scale for Handwriting.¹⁴

Kansas City Scale of Handwriting.¹⁵

¹⁴ W. S. Benson & Co., Chicago.

¹⁵ Kansas City School Department, Kansas City, Mo.

Learner's Diagnostic Practice Sentences in Handwriting.¹⁶
Lister-Myers Handwriting Scales.¹⁷
Pressey Chart for Diagnosis of Illegibility in Handwriting.¹⁸
Starch-Wise Scale for Measuring Handwriting.¹⁸
West Chart for Diagnosing Elements of Handwriting.¹⁶
Zaner Handwriting Scales.¹⁹

SELECTED REFERENCES

Ayres, L. P., "A Scale for Measuring the Quality of Handwriting of School Children" (*New York Bulletin* No. 113, Russell Sage Foundation).

Freeman, F. N., *The Teaching of Handwriting* (Boston, Houghton Mifflin Company, 1914).

Freeman, F. N., "Handwriting Movement" (*Supplementary Educational Monographs*, Vol. II, No. 3).

Freeman, F. N., "Handwriting" (*Fourteenth Yearbook*, National Society for the Study of Education, 1915).

Freeman, F. N., "Practical Studies in Handwriting" (*Elementary School Teacher*, Vol. 14, pp. 167-179).

Freeman, F. N. and Dougherty, M. L., *How to Teach Handwriting* (Boston, Houghton, Mifflin Company, 1923).

Freeman, F. N., "Handwriting" (*Third Yearbook*, Department of Superintendence, National Education Association, pp. 205-216, 1925).

Gray, C. T., "A Score Card for the Measurement of Handwriting" (*Bulletin of the University of Texas*, No. 37, 1915).

Kirk, J. G., "Handwriting Survey to Determine Grade Standards" (*Journal of Educational Research*, Vol. 13, pp. 181-188, 259-272 March and April, 1926).

Koos, L. V., "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools" (*Elementary School Journal*, Vol. 18, 1925, pp. 423-446).

¹⁶ Public School Publishing Co., Bloomington, Ill.

¹⁷ Brooklyn Training School for Teachers, Brooklyn.

¹⁸ D. Starch, 1374 Massachusetts Ave., Cambridge, Mass.

¹⁹ C. A. Gregory Co., 345 Calhoun St., Cincinnati.

Starch, D., "The Measurement of Efficiency in Handwriting" (*Journal of Educational Psychology*, Vol. 6, 1915, pp. 106-114).

Starch, D., *Educational Psychology*, Chapter XVIII (Revised) (New York, The Macmillan Company, 1927).

Thorndike, E. L., "Handwriting" (*Teachers College Record*, March, 1910).

Thorndike, E. L., "The Measurement of Ability in Handwriting" (*Teachers College Record*, November, 1914).

CHAPTER VII

READING

Importance and problems of reading.—Reading, the most fundamental subject of the elementary school curriculum, presents many interesting and difficult psychological problems. Very early in its life, the child learns to associate certain sounds with certain objects as a result of ordinary experiences. As a result of further experiences and observations the child learns to incorporate words into spoken language. It learns to express its wants, feelings, and ideas by the use of words and sentences. This process is well developed in the child before there is any attempt to teach it to read. Reading for the beginner is the process of associating written or printed symbols with certain sounds. As taught by modern methods, these sounds are the words the meaning of which the child already knows. The problem for the child and the teacher is, therefore, a double one: that of associating symbols and words with their spoken sounds, and also that of getting meaning from the printed or written words. The first is a mechanical process and requires long, patient drill. The second, although dependent upon the first, is ultimately by far the more important, since the purpose of reading is the gaining of meaning from the printed page.

In the beginning, the words must be spoken by the child. This is the period of oral reading. Later the

words may and for the most part should be read silently by the child. Here the mechanics drop into the background, and reading becomes a process of thought comprehension. This thought-getting may consist in the mere understanding of the meaning of the printed words; it may consist in the understanding of sentences or in the interpretation of paragraphs or whole selections; or it may be a combination of all these factors.

In emphasizing the importance of reading it is pointed out that more than one-fourth of the time in our elementary schools is devoted to the formal study of reading. But much more significant than this is the fact that it is by means of reading that the pupil, as well as the adult, obtains the larger share of his information. It is very evident that the pupil's success in history, geography, literature, hygiene, and other subjects depends very largely on his ability to extract meaning from the printed page. Too often the teacher does little more than quiz the pupil on his reading; but at the best she only explains, elaborates, and interprets the lesson assignment. Even in a subject like arithmetic, reading is a very important factor. P. W. Terry¹ has shown that much of the pupil's difficulty in arithmetic problems lies in poor reading. Manifestly a pupil can not solve a problem if he does not understand what the problem means; nor can he solve problems rapidly if his reading is laborious.

What has been said of reading in the grades applies even more emphatically to the high school and college

¹ "How Numerals Are Read," *Supplementary Educational Monographs*, No. 18 (Chicago University Press, 1922).

student. Many of the failures in high school can be traced directly to poor reading habits rather than to a lack of intelligence. With the great increase in the amount of reading demanded of the high school pupil comes the added demand for rapid, efficient, silent reading. If such reading habits have been developed, the pupil is prepared for his tasks. If he be poorly prepared in the mechanics of reading, his chance of success is limited. It is the verdict of a large number of high school teachers that many pupils are not well prepared. In fact, relatively few are as well prepared as they should be. This condition may be due either (1) to an inability to comprehend the printed page, (2) the formation of habits of slow reading, or (3) both of these combined.

By methods too commonly used in teaching children to read, thought-getting has not been emphasized as the principal aim in reading. The child has been taught that word pronouncing is the end. This has sometimes come from improper training in phonics and prolonged emphasis on oral reading. The result is that the pupil does not develop the habit of reading for meaning. Not only has comprehension of the printed page been neglected, but too often rate of reading has been retarded through emphasis upon oral habits. It has been found that the major period of growth in the mechanics of reading, as determined by eye movements, takes place within the first four grades.² The significance of this fact, of course, is that there must be efficient teaching

² Buswell, G. T., "Fundamental Reading Habits: A Study of Their Development," *Supplementary Educational Monographs*, No. 21 (University of Chicago, 1922).

of reading in the early grades if the mechanics of reading are to be adequate and effective, so that later instruction may be focused chiefly upon comprehension and interpretation.

Speed of reading is seldom sufficiently emphasized in school instruction. In fact, pupils are often warned against reading too rapidly. This caution may be justified in oral reading when the pupil attempts to proceed too fast to permit proper articulation. But there is little danger in rapid silent reading. In fact, most persons could read considerably faster than they do without loss in comprehension. An investigation was made to study the effect of reading at normal, rapid, and slow speeds on ability to recall what was read.³ It was found that there was very little difference in the amount recalled by any of the three methods of reading. When the gain in time is taken into consideration, reading at the most rapid rate was much more efficient. The elementary school pupils gained about one-fourth by the rapid reading, while the high school and college students gained even more.

In this connection it is worth noting that most individuals by a little systematic practice could materially increase their rate of reading. E. B. Huey,⁴ after enumerating a number of experiments in which speed of reading was greatly increased without loss in comprehension, reports that he doubled his own reading rate as a result of practice in rapid silent reading.⁵

³ Gilliland, A. R., "The Effect of Rate of Silent Reading on Ability to Recall" (*Journal of Educational Psychology*, November, 1920, Vol. XI, p. 474 ff.).

⁴ Huey, E. B., *The Psychology and Pedagogy of Reading*, p. 180.

⁵ It sometimes happens that a pupil is a genuine "non-reader" and cannot learn to read except through the use of special methods of instruc-

When comprehension is poor and the rate of reading slow, a pupil is seriously handicapped, and it is only by prolonged and concentrated effort that even a mediocre amount can be accomplished. If this effort is not forthcoming, the pupil's chances of failure are increased. The question of the relationship between rate of reading and comprehension has been of marked interest to investigators, and it is one concerning which there is some misapprehension. It is not true that the rapid reader in general remembers little of what he has read, whereas the slow reader is superior in comprehension, for it has been demonstrated that high rate and good understanding are related, and that low rate and poor understanding are related. This fact is of great importance to the teacher.

Types of tests.—Tests of reading may serve two general purposes: (1) to measure the pupils' progress in rate and comprehension during the years when they are receiving instruction in reading, and by means of diagnostic tests to detect defects so that corrective measures may be applied; (2) as an instrument to detect in later grades whether a pupil's poor work may not be due, in part at least, to poor reading.⁶ For the latter purpose, tests of general comprehension and speed should first be used.

tion. When a child, apparently bright and intelligent in most matters, is found to be unable to learn to read, he is a case for the oculist first, and then for the specialist. For a discussion of this problem of non-readers, see "Teaching Reading to Non-Readers," by Dearborn, W. F., *Elementary School Journal*, December, 1929, pp. 266-269. Also "Special Disabilities in Learning to Read and Write," by Lord, E. E., Carmichael, L., and Dearborn, W. F., *Harvard Studies in Educational Psychology and Educational Measurements*, Vol. 2, No. 1, 1925.

⁶ Judd, C. H., and others, "Reading, Its Nature and Development," *Supplementary Educational Monographs* (University of Chicago, Chapters V and VI).

Many tests have been constructed for measuring ability in reading. One of the best known of these is designed to measure ability in the mechanics of oral reading. Others measure ability to understand the meaning of words, sentences, or whole selections. Various methods of recording responses are used. In some of the tests the pupil is required to draw a line under or around certain words to indicate the correct response to a question on the passage read; in others he is called upon to reproduce the story orally or in writing. In still others the pupil is to answer in very few words certain questions based upon what has been read.

The teacher may become perplexed by the large number of tests and might reasonably ask which of them to use. The answer depends largely upon what aspect of the reading process is to be tested. If the teacher desires to find out whether the pupils are prepared in the mechanics of oral reading, she will use the Gray Standardized Oral Reading Tests. If she is concerned with the ability of the pupils to understand the meaning of words, the Pressey or Thorndike Visual Vocabulary Scale may be used, the Pressey Test for the lower grades and the Thorndike in grades three to second year high school. If the teacher desires to eliminate the influence of handwriting as a factor in the measurement of reading comprehension, the Kansas Silent Reading Test, the Burgess Scale for Measuring Ability in Silent Reading, the Courtis Silent Reading Tests, or the Monroe Standardized Reading Test may be used. Some of these are much more comprehensive than others. Some emphasize thought-getting while others re-

quire interpretation and reasoning. The Gray Silent Reading Tests require the pupils to reproduce the story and answer a list of questions based on the story. The Haggerty tests measure word meaning, sentence meaning, and paragraph meaning. The teacher, therefore, must decide what aspect of the reading process is to be examined, and she must select the test accordingly.

GRAY STANDARDIZED ORAL READING PARAGRAPHS AND THE ORAL READING CHECK TESTS

Description of the test.—This oral reading test devised by W. S. Gray, consists of twelve short paragraphs ranging in difficulty from very easy material in the first paragraph for pupils of the lower grades to paragraphs difficult enough to tax the ability of high school pupils. This increased difficulty consists largely in the use of longer and more unusual words. The purpose of the test is to measure ability in the mechanics of oral reading. It is an individual test.

As the pupil reads one after another of the paragraphs the examiner measures the time required to read each paragraph and records the errors in reading. Six types of errors are recorded. These errors are:

1. Gross Errors—Total mispronunciation of words.
2. Minor Errors—Partial mispronunciation, as wrong vowel sounds or accent.
3. Omissions—Leaving out a word in the reading.
4. Substitutions—Reading another word instead of the one in the text.
5. Insertions—Adding words not in the text.

6. Repetition—The rereading of two or more words after they have already been read.

In order to facilitate the noting of errors, a method of recording is suggested, as in the following illustration.

The sun pierced into my ^{many} large windows. It was the opening of October, and the sky was of a dazzling blue. I looked out of my window and down the street. The white houses of the long, straight street were almost painful to the eyes. The clear atmosphere allowed full play to the sun's brightness.

"If a word is wholly mispronounced, underline it, as in the case of 'atmosphere.' If a portion of a word is mispronounced, mark appropriately, as indicated above: 'pierced' pronounced in two syllables, sound long a in 'dazzling,' omitting the s in 'houses' or the al from 'almost' or the r in 'straight.' Omitted words are marked as in the case of 'of' and 'and'; substitutions as in the case of 'many' for 'my'; insertions as in the case of 'clear'; and repetitions as in the case of 'to the sun's.' Two or more words should be repeated to count as a repetition."

"Each pupil should be allowed to continue reading until he makes at least the following number of errors in each of two paragraphs: 5 errors or more in 40 or more seconds, or 7 or more errors in case the paragraph is read in less than 40 seconds."

Scoring the test.—A pupil's score is a combination of his rate of reading and the number of errors made. A table is provided on the score sheet for determining the combined score. The score sheet also presents a

somewhat elaborate method of obtaining the final individual or class score.

The Gray Oral Reading Test suffers from the fact that the scoring can not under the circumstances be as objective as is desirable, inasmuch as the child's responses are not automatically recorded, so that they might be judged by any qualified person.

SAMPLE PARAGRAPHS FROM THE GRAY STANDARDIZED
READING TESTS

I

A boy had a dog.

The dog ran into the woods.

The boy ran after the dog.

He wanted the dog to go home.

But the dog would not go home.

The little boy said,

“I cannot go home without my dog.”

Then the boy began to cry.

6

The part of farming enjoyed most by a boy is the making of maple sugar. It is better than blackberrying and almost as good as fishing. One reason why a boy likes this work is that someone else does most of it. It is a sort of work in which he can appear to be very industrious and yet do but little.

12

The hypotheses concerning physical phenomena formulated by the early philosophers proved to be inconsistent and in general not universally applicable. Before relatively accurate principles could be established, physicists, mathematicians, and statisticians had to combine forces and work arduously.

NORMS FOR GRAY STANDARDIZED READING PARAGRAPHS

<i>School grade</i>	1	2	3	4	5	6	7	8
<i>Grade score</i>	31	42	46	47	48	49	47	48

A fifth grade child therefore should make a score of 48 on the test. The apparent lack of improvement in the score for the different school grades is due to the different credits for the first paragraph.

Function of the test.—This is a very useful reading test. While it does not measure all the factors in oral reading, such for example as expression or thought-getting, the factors which are measured are fairly definite. The test is diagnostic in the sense that the records show not only the number but also the types of errors made in oral reading. The teacher by reference to the score sheet can determine the factors in which her pupils are below average in their reading. If only a few pupils in the class are deficient, they should receive individual and special drill intended to overcome their particular deficiencies. If, on the other hand, the class in general is suffering from one or several types of errors, the teacher should, of course, focus her attention on those points in order to eliminate them. If the class is uniformly poor, the causes should be determined. It may be that the pupils have had poor preparation; or perhaps they are of low grade mentality; or it is possible that the time being devoted to instruction in reading is inadequate. It is also possible that the method being employed is faulty. Each of these possibilities should receive careful consideration.

The Oral Reading Check Tests.—The Gray Oral Reading Paragraphs may be supplemented by the Oral

Reading Check Tests. The purpose of the latter is “(1) to secure accurate measures at frequent intervals of the progress of pupils in rate and accuracy of oral reading, and (2) to secure detailed information which will aid in determining the specific nature of the difficulties which poor readers encounter.”

The check tests consist of passages at four levels of difficulty: the first for grade 1, the second for grades 2 and 3, the third for grades 4 and 5, and the fourth for grades 6, 7, and 8. At each level there are five tests approximately equal in difficulty. The first test at the appropriate level is given; and the remaining tests of that level are administered at intervals of two, three, or four weeks. During the course of testing, progress in both rate and accuracy are noted. Norms are provided for rate and accuracy for each grade which may be used in studying the progress of the pupil. The check test is also accompanied by individual record sheets on which may be recorded the results “of a progressive analysis of errors in oral reading.” The errors include such types as mispronunciations, enunciation, substitution, omissions, insertions, repetitions. The check tests and the record sheets provide the teacher or examiner with the means for carrying out a systematic study of difficulties and improvement in oral reading.

PRESSEY FIRST GRADE WORD READING TEST AND FIRST GRADE READING SCALE

Description of the tests.—The First Grade Word Reading Test is made up of three parts, the first two

being distinctly word recognition tests, and the third in the nature of a vocabulary test. Each part has 15 rows of words, of five words each. In the first two parts the examiner merely reads aloud the word to be identified, and the pupils are required to draw a circle around that word. In part three the examiner reads the definition of one word in each row, and the pupil is to draw a circle around the word defined.

The words of this test were taken systematically from the Gates Primary Word List which is made up of 1,500 words commonly appearing in primary reading material. The Pressey test is devised so as to sample the 1,500 words in each of the three sections.

FROM THE PRESSEY FIRST GRADE READING SCALE

1. is the you a said
2. do we come are ball
3. baby one with that have
4. good was on this his
5. mouse has your match bird
6. oh give for mother ran
7. fly very water as milk
8. home help blow some girl
9. three won't be name will
10. rabbit bumblebees over shall bear

The First Grade Reading Scale is composed of a word test and a sentence test. The word test consists of twenty-five rows of five words each. As in the word test above described, the pupils are directed to draw a line around a certain word in each column. Each succeeding list of words is somewhat more difficult than the preceding. The sentence test is similar except that the

fifteen rows are composed of short sentences instead of disjointed words. There is no time limit for these tests. Alternate forms of both the Vocabulary and Reading Scale are available.

Function of the tests.—The general purposes of these tests are very similar. They constitute measures of recognition of words in the reading vocabulary of first grade children, thereby filling the need of a measure of the knowledge of words out of context for the first grade⁷ very much the same as the Thorndike Visual Vocabulary Scale does for the upper grades.

THORNDIKE VISUAL VOCABULARY SCALE AND THE TEST OF WORD KNOWLEDGE

Description of the scale.—The scale consists of a graded series of words which the pupil is to classify according to certain specified groups. This classification is accomplished by placing a letter or word under each word. For example, the letter "F" is to be written under every word that means flower, the letter "A" under every word that means animal, and the word "Bad" under every word that means something bad to be or do.

The words are arranged in groups, and these groups are assigned numerical values on the basis of their difficulty. The first group in scale A contains five words with a value of 4. The last group contains three words with a value of 11. A preliminary test, sometimes called a shock absorber, precedes the test proper to familiarize the pupil with the nature of the test. There are four

⁷ A reading test containing a measure of vocabulary also constitutes a part of the Pressey Second Grade Attainment Scale.

forms of the scale: A2z, A2y, Bx, and By. The four scales are of approximately equal difficulty and may be used from grade three to second year high school.

Scoring the scale.—After the preliminary test has been given the child is furnished a copy of the scale. The directions are printed on the scale. No time limit is set; hence, the child is allowed to work until he finishes. The pupil's score is the number of the highest numbered groups of words in which he makes not more than a single error. The numbers roughly indicate the average effect of this number of years of training on the vocabulary of the child.

Function of the scale.—This scale is especially valuable as a measure of the pupil's knowledge of the meaning of words out of context. While word knowledge is really not a part of the reading process proper, it is such an essential basis for success in reading that the Thorndike Scale is here listed as a reading test. If a pupil is doing poorly in his reading, the teacher may well use such a test as this to determine whether or not the pupil's difficulty is with vocabulary. If so, the pupil should be drilled on word meanings. If the pupil makes a good score in the Visual Vocabulary Scale and still is a poor reader, the tests that measure the mechanics of reading or sentences and paragraph meanings should be used in order to locate his difficulty.

FIRST HALF OF THE THORNDIKE READING SCALE D (VISUAL VOCABULARY)

Write the letter W under every word that means something about *war* or *fighting*.

Write the letter B under every word that means something about *business* or *money*.

Write the letters CHU under every word that means something about *church* or *religion*.

Write the letter R under every word like *father* or *wife* that means something about *relatives* or the *family*.

Write the letters COL under every word that means a *color*.

Write the letter T under every word like *now* or *then* that means something to do with *time*.

Write the letter D under every word like *here* or *north* that means something about *distance* or *direction* or *location*.

Write the letter N under every word like *ten* or *much* that means something about *number* or *quantity*.

4x. camp, flag, west, mother, two, general, green, troops, south, fort

4½x. gray, cousin, pink, uncle, yellow, hour, pay, aunt, early, commander

5x. marriage, defeat, many, afternoon, guard, buy, captive, military, relation, late

6x hymn, defend, across, merchant, noon, forty, conquer, dagger, profit, tuesday

6½x. month, dozen, fortress, cavalry, tax, bishop, below, october, million, owe

7x. fortification, ownership, there, year, june, half, scarlet, soon, november, beneath

Thorndike Test of Word Knowledge.—This test, of which there are four equivalent forms, is made up of 100 rows of words. In each row the first word is the word on the meaning of which the pupils are to be tested. Following it, in the same row, are five other words, one of which is to be selected as meaning the same, or most nearly the same, as the first. The words are, of course, of increasing difficulty.

The test is satisfactory in grades 5 to 10, although it may be used in grade 4 and with high school students beyond the tenth grade. Norms, perhaps somewhat

high because based upon a selected group of cities, are provided for grades 4-9.

Inasmuch as this test is based upon previous investigations which have yielded 10,000 English words

1. afraid	1 full of fear....2 possible ...3 necessary .. 4 raid.....5 ill1
2 baby	1 manner .2 trembling . 3 young child ..4 notice....5 soft 2
3. divide	1 mount....2 pound... 3 hold . 4 cut into parts .. 5 add together3
4. require	1 revenge...2 report.. .3 need....4 reward....5 return4
5. action	1 play. . 2 deed....3 mention .. 4 opinion. ..5 crime5
96. insecure	1 vault,...2 abnormal....3 unsafe .. 4 extant....5 insure96
97. madrigal	1 song....2 mountebank....3 lunatic....4 ribald....5 sycophant97
98. nauseous	1 boorish ...2 loathsome....3 synchronous .4 seafaring....5 inopportune98
99. pact	1 puissance. .2 remonstrance. .3 agreement ...4 skillet... 5 pressure	.. 99
100. distend	1 swell....2 prevent... 3 hoodwink . 4 put an end to... .5 inaugurate100

FROM THE THORNDIKE TEST OF WORD KNOWLEDGE (FORM A)

with a measure of the importance of each,⁸ it is valuable as a measure of an individual's necessary word knowledge. The test, however, is what may be called a measure of one's "passive" vocabulary; that is, it examines one's ability to identify words, but it does not examine ability to *use* words in language, whether oral or written. The latter type of vocabulary is known as "active."

MONROE STANDARDIZED SILENT READING TEST

(Revised)

Description of the test.—This test, devised by W. S. Monroe, measures both speed and comprehension. The reading material consists of a series of short paragraphs, each paragraph followed by a list of words. The pupil is directed to underline one of these words on the basis of information contained within the paragraph. Four minutes is the time allowed in which the pupil is to read and underline as many words as he can. Test I is for grades 3, 4, and 5 and Test II is for grades 6, 7, and 8. There are three forms of equal difficulty for each test.

TWO PARAGRAPHS FROM THE MONROE STANDARDIZED SILENT READING TEST. REVISED FROM TEST II FOR GRADES 6, 7, AND 8, FORM 3

537 11. Beside our house was a little hut where a holy man lived
550 in charge of an adjoining shrine, earning money for himself
562 and for the shrine by polishing little pieces of marble as
 mementos for visitors.

⁸ Thorndike, E. L., "Word Knowledge in the Elementary School," *Teachers College Record*, September, 1921; also "The Teacher's Word Book."

573 Draw a line under the word which best describes this holy man.
industrious good foolish lazy sad

590 12. Nanook, once so full of life, now knew perfectly well
 602 that it was all over with him. Head and tail down, the picture
 615 of resigned dejection, he stood like a petrified dog.

622 Draw a line under the word which best describes the dog
 Nanook.
 634 *angry frightened active hungry down-hearted*

Scoring the test.—The comprehension score is the number of exercises the pupil answers correctly in the four minutes allowed for the test. This score may then be transferred into an Accomplishment Age score by means of a table given in the Teachers' Handbook. The rate of reading may also be transferred into an Accomplishment Age score by reference to the same table. The author suggests that these two accomplishment scores may be averaged as a single measure of silent reading ability.

The following norms are given for both rate and comprehension:

GRADE MEDIAN MONROE SILENT READING REVISED

Grade	Comprehension			Rate		
	Form 1	Form 2	Form 3	Form 1	Form 2	Form 3
III	3.8	3.8	3.8	82	78	81
IV	7.7	7.7	7.7	122	116	121
V	9.8	9.8	9.8	142	135	141
VI	11.0	11.1	11.7	159	164	179
VII	12.5	12.6	13.3	171	176	192
VIII	13.7	13.8	14.6	185	191	208

These standards are derived from scores which for the most part represent the achievements of pupils early in the school year. The returns are about equally distributed between rural schools and city schools.

Function of the test.—This test is easy to give and requires only a small amount of time. It requires little writing on the part of the pupil and the grading is not difficult. As has been pointed out in another place, a test should, so far as it can, isolate the factor or factors to be measured, and measure them independently of others as far as possible. In the case of reading, both rate and comprehension can be measured at the same time. But in this test the pupil's answer depends upon reasoning ability as well as reading ability. In so far as this is true it becomes in part a test of intelligence rather than one of reading ability. The method of measuring rate of reading may also be questioned. Rate is measured by the amount of work completed in four minutes. Part of this time is devoted to reading and another part to underlining the answers. A pupil might read very rapidly and yet take a long time in the underlining, and thereby receive a spuriously low score in rate. Monroe's method of averaging comprehension score and rate score as a measure of reading ability is also open to question.

THORNDIKE-MCCALL READING SCALE

Description of the scale.—This silent reading scale is made up of a series of paragraphs each followed by a list of questions. There is a preliminary paragraph for practice and nine other paragraphs ranging in difficulty from the first, which is easy enough for a second grade pupil, to one difficult enough for a high school senior. There are thirty-five questions based upon the paragraphs. Thirty minutes are allowed for the test,

and during this time the pupil may read and reread the paragraphs as many times as he likes.

Scoring the scale.—The papers are graded by determining the number of questions answered correctly. This number is translated into what is called a T score⁹ by means of a table which is furnished with the scale. Ten equivalent and interchangeable forms of the scale have been constructed. The following norms are given:

GRADE NORMS

<i>At end of</i>	<i>Norm</i>	<i>Approx. No. Pupils</i>	<i>At end of</i>	<i>Norm</i>	<i>Approx. No. Pupils</i>
2A	26	200	8A	59.6	5,000
2B	30	300	8B	60.9	10,000
3A	33.7	3,000	9A	61.5	
3B	37.3	5,000	9B	62.1	1,000
4A	39.6	5,000	10A	62.9	
4B	41.8	10,000	10B	63.6	1,000
5A	44.9	5,000	11A	64.5	
5B	48.0	10,000	11B	65.4	1,000
6A	50.9	5,000	12A	66.8	
6B	53.7	1,000	12B	68.1	1,000
7A	56.0	5,000	Superior		
7B	58.3	10,000	Teachers	72.0	300

⁹ The T-unit equals 1/10 of the S. D. (standard deviation) of a distribution for 12 year old pupils. See Chapter XVIII for a definition and explanation of the S. D.

For a discussion of the T scale see McCall, Wm. A., *How to Measure in Education*, (New York, The Macmillan Company, 1922, Chap. X, pp. 272-308).

Sample Paragraph from the
THORNDIKE-MCCALL READING SCALE FOR THE UNDERSTANDING
OF SENTENCES—FORM 4

Write your name here.....
School..... Grade..... Date.....
How old are you?..... When is your birthday?.....

This is to be a reading contest. You will read paragraphs like this one, and answer questions like those you see below. Answer every question you can. If you come to a question you can't answer skip it and go on. Go back to it later. If you finish before you are told to stop, go back and make sure you have made no mistakes. When possible, the answers to the questions must be found in the paragraph. You may read the paragraph as many times as you need to. You will have enough time but don't waste it. Play fair. Don't look at anyone else's paper. You will be told your score later.

Read this and then write the answers. Read it again if you need to

In August, Arthur and his Cousin Kate went in the train to visit their grandfather, Mr. Peters, at Oak Farm. They played in the brook, picked blackberries, and hunted for eggs in the barn. They played with Bob Peters and Nan Allen. Bob was nine years old; Nan was eleven.

5. How old was Nan?.....
6. Which was older, Bob or Nan?.....
7. Does the story tell how old Kate was?.....
8. Does the story say that Bob and Nan went in the train?.....
-

Function of the scale.—The subject-matter of this test is the ordinary sort of material that a child might be expected to read. The answers to the questions are based directly on the material contained in the paragraphs. Not much writing is required of the pupils and

yet the questions are not of the "yes" and "no" type. The test does not measure rate of reading except very indirectly, and should not be used for that purpose. The alternative forms of the test make it possible to make frequent retests of pupils in order to measure progress. In general, this is a very satisfactory test of comprehension of silent reading.

BURGESS SCALE FOR MEASURING ABILITY IN SILENT READING

Description of the scale.—The Burgess Scale consists of a series of twenty pictures each followed by directions. The pupils are told that each paragraph "tells them to do something to the picture above it with their pencils." They must read carefully to make sure what they are to do. The paragraphs are to be read and marked in order, starting at the top and working down. The pupil must do as many as he can in the five minutes allowed for the test.

Scoring the scale.—Every paragraph is counted as correct in which the marking of the picture, no matter how crude it may be, exactly follows instructions. The pupil's score is the number of paragraphs marked correctly. This score may be translated into a score on the basis of 0 to 100 for any school grade by reference to the accompanying table.

Test Score	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Grade	8	...	0	26	32	38	44	50	56	62	68	74	80	86	92	98	100
"	4	..	0	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100
"	5	..	0	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100	..
"	6	..	0	2	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100
"	7	..	0	2	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100
"	8	..	0	2	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100

The scores on this table are so constructed that the score of the "average" pupil of any grade is 50. Above



1. This naughty dog likes to steal bones. When he steals one he hides it where no other dog can find it. He has just stolen two bones, and you must take your pencil and make two short, straight lines, to show where they are lying on the ground near the dog. Draw them as quickly as you can, and then go on.



5. Have you ever seen such a strange bird? He is hard to find because he sleeps in the woods during the day and does not come out until night. Take a pencil and tell people what the bird's name is by writing the word OWL, with a capital O, under the books on which the bird is standing.



2. This man is an Eskimo who lives in the far north where it is cold. There has just been a big storm, and all the ground is white with snow. The man has been walking and has made many footprints in it. With your pencil quickly make four of these in the snow just behind him.



6. This small chap is afraid to start for school. The teacher will scold unless he brings his books, but the big owl is sitting on them. Grasp your pencil bravely and cross the owl out of the previous picture with two black lines, so that the child can rescue his belongings. Remember not to use more than just two lines.



3. This book is lying on the desk, but it is hard to make it stay open. With your pencil draw a single straight line to represent a ruler lying across the book to hold the pages open. Be sure to make the line from one side to the other, across the book, instead of making it go up and down.



7. These two flags are used as signals to give notice of changes in the weather. The white flag means fair; so you may now take your pencil and make a capital F under the white flag, to stand for fair. The blue flag means storm, so make a capital S under the blue one.

SAMPLE PARAGRAPHS FROM THE BURGESS SILENT READING SCALE

average is denoted by a score of more than 50 and below average by a score below that mark. These scores are for February 1st. For other months of the school year specific additions or subtractions are to be made, as indicated in the author's directions.

Function of the scale.—The author of this scale states that it was so devised that it should be free from four fundamental limitations of the ordinary scale. (1) It is expected that this scale will measure ability in reading and not other abilities such as handwriting

or English composition. (2) The scale is uniform and relatively equal in difficulty throughout. It measures one and only one ability. (3) It is easy to administer and score. (4) The score for each school grade is available for comparison with the standing of other children.

In so far as this test conforms to these four principles it is superior to some of the other scales. But a number of the paragraphs call for slight drawing ability which might conceivably be a handicap to some children.

This scale does not measure rate of reading and should not be employed for that purpose. Inasmuch as it measures only comprehension, it lacks the diagnostic value of those which separate the two factors.

COURTIS SILENT READING TEST

Description of the test.—Part I of this test consists of a story of two ordinary pages in length. The pupil is directed to read silently and is given three minutes to read as much as he can. At half-minute intervals, on a signal from the examiner, the pupil puts a mark around the last word read. He then proceeds to Part II which consists of the same story that has just been read, but here it is broken up into fourteen short paragraphs. A series of five questions to be answered by "yes" or "no" follow each paragraph. There is a preliminary paragraph as an example, after which the pupil is given five minutes to answer as many questions as he can, in order. In the second section the examiner has the pupil make a mark every minute. He is allowed

to refer to the paragraphs for the answers. The test is constructed for use with grades 2 to 6.

Scoring the test.—The comprehension score is the number of questions answered correctly minus the number answered incorrectly, divided by the number answered correctly. For example, if the pupil answers 28 questions correctly but gives a wrong answer to 12 other questions, his score is 16 divided by 28 or 57%. This score is called the Index of Comprehension. The norms for the different school grades are as follows:

COURTIS SILENT READING TEST NORMS

<i>School Grade</i>	2	3	4	5	6
<i>Index of Comprehension score</i>	59	78	89	93	95

Norms are also given for the number of words read per minute. This rate is here presented together with the norms given by Starch and W. S. Gray, for purposes of comparison.

WORDS READ PER MINUTE

<i>School Grade</i>	2	3	4	5	6	7	8
Starch	108	126	144	168	192	216	210
Gray	90	138	180	204	216	228	240
Courtis	84	113	145	168	191		

The differences in these norms are no doubt to be accounted for in part by the difference in the nature of the subject-matter read.

Function of the test.—The subject-matter of the Courtis Test is unusually interesting, a factor that is lacking in some of the other reading tests. The test is easily scored, as all answers are either right or wrong. Some criticism may be made to questions answered

either "yes" or "no" even though an attempt is made to eliminate the effects of guessing by what is known as the "right-minus-wrong" method of scoring. By this method of scoring, an individual who merely guesses would probably get a score of zero, inasmuch as the chances are even that his guess will be right or wrong. At least that is the assumption. But there are several serious psychological objections to the "right-minus-wrong" procedure. One of these is the fact that the child who works very slowly but carefully is likely to receive a score on comprehension as high as or higher than that of the more rapid worker, who might be penalized by his increased errors. A further objection is that the "R-W" procedure assumes that there is just as much likelihood of a correct or an incorrect guess when the pupil is not certain. This places the matter on a purely and unjustifiable chance basis.

The author of the test states that it measures "Simply the ability to read silently and understand a simple story and simple questions about the story."

GRAY SILENT READING TEST

Description of the test.—The reading material for this test consists of three short stories. One story, "Tiny Tad," is for use in the second and third grades; another, "The Grasshopper," is for the fourth, fifth, and sixth grades; the other story, "Ancient Ships," is for use in the seventh and eighth grades. This is an individual test and both rate and comprehension are scored.

Scoring the test.—Rate of reading is measured by the time required to read a section of the story under

natural or as near natural conditions as possible, for the child is not aware of the fact that he is being timed. The examiner watches the child closely and starts his stop-watch when the child raises his eyes to read the second section of the story. He stops the watch when the pupil looks up to begin reading the third section.

The comprehension score is the average of two scores: (1) the number of words used by the pupil in the reproduction of the story, after discounting repetitions and wrong or irrelevant statements, and (2) answers to a list of questions based upon the story. Standards for rate and comprehension follow:

THE GRAY SILENT READING TEST NORMS

<i>School Grade . . .</i>	2	3	4	5	6	7	8
<i>Quality score . . .</i>	32	37	29	32	39	22	27
<i>Words read per</i>							
<i>second . . .</i>	1.50	2.30	2.20	2.57	2.74	2.69	2.87

Function of the test.—This test is valuable in that (1) it presents interesting reading material to the pupil; (2) its method of measuring understanding is comprehensive; and (3) it measures rate under normal conditions better than many other tests. Its disadvantages are that the scoring is not altogether objective, and the difference in reading material from grade to grade makes it impossible to measure progress continuously over the grades. It has the practical though less serious disadvantage of requiring too much time, inasmuch as it is an individual test.

HAGGERTY READING EXAMINATION

Description of the examination.—The Haggerty Examination at present consists of two parts: Part One,

8. Put a cross over each bird that is on the ground.



9. Put a cross on each child that is hiding.
10. Put two lines under the girl who has found the children.



Once a hungry wolf was about to eat a poor little pig.
The little pig jumped into a big kettle and saved herself just in time.

11. Put a line under the animal which was about to eat the pig.
12. Put a cross under the place where the pig hid.

called Sigma 1, is for grades 1 to 3, and Sigma 3 for grades 6 to 12. Sigma 1 is made up of a preliminary exercise and two tests. Test 1 contains sets of pictures, several paragraphs chiefly descriptive of the pictures, and twenty-five tasks based on the pictures and paragraphs. The score on this test is the number of items answered correctly. Test 2 contains twenty questions, each followed by "no," "yes," the correct answer to be underlined. Twenty minutes are allowed for the examination.

Sigma 3 is composed of three preliminary tests and three regular tests. Test 1 is a vocabulary test consisting of fifty words, each followed by four or more words, one of which is a synonym or definition. The pupil is to underline the best definition for as many of the fifty words as he can. Test 2, sentence meaning, consists of forty statements followed by "yes," "no." Test 3, paragraph reading, is composed of a series of seven paragraphs, each being followed by true and false statements. Correct statements are to be underlined. Twenty minutes are allowed for this examination.

Scoring the examination.—The score for all these tests is the number of questions answered correctly. In Sigma 1, test 1 and test 2 are scored separately. Grade norms based upon the results for 6,000 children are given as follows:

NORMS FOR THE HAGGERTY READING EXAMINATION—
SIGMA 1

<i>School Grade</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
Score Test 1	4	12	16	20
Score Test 2	2	8	14	18

In Sigma 3 the scores for all three tests are combined. Norms are given as follows:

NORMS FOR THE HAGGERTY READING EXAMINATION—
SIGMA 3

School Grade	5	6	7	8	9	10	11	12
Score	40	54	68	80	93	104	112	118

Function of the examination.—The Haggerty Examinations are interesting to the pupils, and they are comprehensive, since they measure word meaning, sentence meaning, and paragraph meaning. The factor of writing is practically eliminated. On the other hand, these examinations are complex and in places measure a mixture of several different aspects of reading ability, in addition to making demands on one's information, some of it very specialized. Consequently parts of the test may almost be designated as "intelligence tests." However, as tests of general reading ability, exclusive of rate, the Haggerty Examinations have much to commend them.

Comparison of reading tests.—Data comparing the merits of current tests in reading are not as numerous as they should be, for it frequently happens that authors of tests fail to supply the necessary statistical facts for the evaluation of their measures. The following table, however, furnishes valuable information with respect to a few of the available tests of reading.¹⁰ It will be noted that with the exception of the Detroit Word test which appears to be inferior, there is little to choose between these measures, as to reliability, for in every

¹⁰ We are indebted to Professor G. M. Ruch for the data contained in this table.

case but one, the reliability (r_{xy}), or consistency of results, is very high.

RELIABILITY COEFFICIENTS, MEANS, AND STANDARD DEVIATIONS

(First, Second, and Third Grade Pupils, with a range in IQ from 87 to 126)

Tests	<i>N</i>	M_x	M_y	SD_x	SD_y	r_{xy}
<i>Group I</i>						
Haggerty Sigma 1	125	11.1	10.9	5.8	6.4	.99 *
Williams	125	12.8	10.3	6.7	5.3	.97
Detroit Word	125	28.3	31.2	10.2	12.8	.68
Stanford Achievement Reading	89	61.1	62.5	30.1	30.0	.94 **
<i>Group II</i>						
Haggerty Sigma 1	77	9.6	9.5	5.4	5.9	.99
Gates Primary	77	84.0	85.2	26.1	27.6	.97
Total of Type 1, 2, 3						

* A correlation of "odds" and "evens" stepped up by the Spearman-Brown prophecy formula. (Probably too large in comparison with other r 's obtained by a more defensible method.)

** The r for the Stanford appears lower than that for Haggerty and Williams. It should be noted, however, that the Stanford was given to a much more homogeneous group (2 grades) than the others (3 grades), thus lowering the r . The data for the Stanford are, therefore, not strictly comparable with the other 3 tests in Group I.

It should be clearly understood that each coefficient of reliability (r_{xy}) applies only to each test singly. The coefficients do not indicate the relationship that exists between the several tests. From the data of the table, it may be said that the several tests (excepting the De-

troit) are highly dependable as to consistency. The scope of some of the tests may at the same time, however, be greater than that of others.

MATERIALS NEEDED

Burgess, May Ayres, A Scale for Measuring Ability in Silent Reading for grades 3 to 8. Four equivalent forms. (Russell Sage Foundation, New York City.) Sample copy 5 cents, \$1.25 per hundred.

Courtis, S. A., Silent Reading Test for grades 2 to 6. (S. A. Courtis, 1807 East Grand Blvd., Detroit, Mich.) \$1.25 for material for testing forty children.

Gray, Wm. S., Oral Reading Test for grades 1 to 8. (Bloomington, Ill., The Public School Publishing Company.) Sample set 5 cents, \$1.00 per hundred. Check tests, \$1.50 per 20.

Gray, Wm. S., Silent Reading Tests, test I for grades 2 and 3, test II for grades 4, 5, and 6, and test III for grades 7 and 8. (Department of Education, University of Chicago, Chicago, Ill.)

Haggerty, J. E., Reading Examinations, Sigma I for grades 1 to 3, Sigma III form A and B for grades 6 to 12, (Yonkers-on-Hudson, N. Y., World Book Company). Sigma I, \$1.00 per 25. Sigma III, \$1.10 per 25; manual of directions for each 25 cents.

Monroe, W. S., Standardized Silent Reading Tests, Revised. Test I for grades 3, 4, and 5, Test II for grades 6, 7, and 8, three forms of each. (Bloomington, Ill., The Public School Publishing Company.) Sample set 10 cents, 80 cents per hundred.

Pressey, S. L., and L. C., First Grade Reading Vocabulary Test. (Bloomington, Ill., The Public School Publishing Company.) Price per package of one hundred, all materials included, 60 cents.

Thorndike, E. L., Visual Vocabulary Scales for grades 3 to 10. Scales A2x, A2y, Bx, and By of approximately equal difficulty. (New York, Bureau of Publications, Teachers College, Columbia University.) Manual of Directions, 40 cents,

stencil 5 cents. Scale 50 cents per hundred. Test of Word Knowledge, \$1.50 per hundred.

Thorndike McCall Reading Scale, for grades 2 to 12. Ten equivalent forms. (New York, Bureau of Publications, Teachers College, Columbia University.) All materials needed supplied with the order, \$2.00 per hundred.

SUPPLEMENTARY LIST OF TESTS

Chapman Unspeeded Reading Comprehension Test.¹¹

Grades 5 and above.

Chapman-Cook Speed of Reading Test.¹¹

Grades 4 to 8.

Dearborn-Westbrook Reading Tests.¹²

Grades 4, 5, and 6.

Detroit Reading Test.¹³

Grades 2 to 9.

Detroit Word Recognition Test.¹³

Grades 1, 2, and 3.

Gates Reading Tests.¹⁴

Primary.

Multiple Skill First Grade Reading Scales.¹⁵

Grades 1B and 1A.

Nelson-Denny Reading Test.¹⁶

Stanford Achievement Test-Reading.¹³

Grades 2 to 9.

Stone Narrative Reading Tests.¹⁷

Grades 4 to 9.

Whipple's High School and College Reading Test.¹⁷

Williams Primary Reading Test.¹⁷

¹¹ J. B. Lippincott, Philadelphia.

¹² W. F. Dearborn, Graduate School of Education, Harvard University.

¹³ World Book Company, Yonkers-on-Hudson, N. Y.

¹⁴ Bureau of Publications, Teachers College, Columbia University.

¹⁵ The Educational Test Bureau, Minneapolis, Minn.

¹⁶ Houghton Mifflin Company, Boston.

¹⁷ The Public School Publishing Company, Bloomington, Ill.

SELECTED REFERENCES

Brooks, F. D., *The Applied Psychology of Reading* (New York, D. Appleton and Company, 1926).

Brooks, F. D., "Reliability of Silent Reading Tests" (*School and Society*, May 31, 1924).

Burgess, M. A., "The Management of Silent Reading" (New York, Russell Sage Foundation, 1921).

Buswell, G. T., "Fundamental Reading Habits: A Study of Their Development." *Supplementary Educational Monographs*, No. 21 (University of Chicago, 1922).

Current, W. F., and Ruch, G. M., "Further Studies on the Reliability of Reading Tests" (*Journal of Educational Psychology*, Sept. 1926), pp. 476-481.

Dearborn, W. F., "The Psychology of Reading" (*Archives of Psychology*, No. 4, 1906).

Dickinson, C. E., "A Study of the Relation of Reading Ability and Scholastic Achievement" (*School Review*, Vol. 33, No. 8, 1925, pp. 616-626).

Foran, T. G., "The Present Status of Silent Reading Tests, Parts 1 and 2." *Catholic University, Educational Research Bulletin*, Vol. 2, nos. 2 and 3 (Catholic Education Press, Washington, D. C., 1927).

Gray, C. T., *Deficiencies in Reading Ability* (D. C. Heath, 1922).

Gray, W. S., "Studies of Elementary School Reading Through Standardized Tests." *Supplementary Educational Monographs*, I, No. 1 (University of Chicago, 1917).

Gray, W. S., "Summary of Investigations Relating to Reading," *Supplementary Educational Monographs*, No. 28 (University of Chicago, 1925).

Huey, E. B., *The Psychology and Pedagogy of Reading* (New York, The Macmillan Company, 1910).

Judd, C. H. and Buswell, G. T., "Silent Reading. A Study of Various Types." *Supplementary Educational Monographs* No. 23 (University of Chicago, 1922).

Judd, et al., "Reading: Its Nature and Development." *Supple-*

mentary *Educational Monographs*, II, No. 4 (University of Chicago, 1918).

National Society for the Study of Education; Report of the Committee on Reading. *24th Yearbook*, Part I, 1925.

O'Brien, J. A., *Reading: Its Psychology and Pedagogy* (New York, The Century Co., 1926).

Reed, H. B., *Psychology of Elementary School Subjects* (Boston, Ginn and Company, 1927).

Sims, V. M., "The Reliability and Validity of Four Types of Vocabulary Tests" (*Journal of Educational Research*, Vol. 20, September, 1929, pp. 91-96).

CHAPTER VIII

ENGLISH LANGUAGE AND COMPOSITION

The problem of measuring language ability.—Thus far we have considered writing, spelling, and reading, all three of which play a part in the use of language. In this chapter we shall present some aspects of the language process and the organization of thought, particularly in writing. Instruction in the English language is concerned in part with (1) the acquisition of words and forms of expression; (2) with the acquisition of grammatically correct English; (3) with the mastery of the mechanical aspects of writing and speaking; and (4) with the acquisition and expression of ideas. The last is particularly important, inasmuch as a large portion of an individual's thinking is conditioned by his control and use of language.

Because of the complexity of factors involved, however, the expression of thought is especially difficult to measure. In addition to the inherent difficulties, judges themselves have not been agreed on the proper standards to be employed in evaluating the written or the spoken word. Yet, by means of scales, more or less objective, teachers and other judges are approaching agreement.¹ It is generally recognized that the ideal

¹ Hudelson, E., "Use of Objective Standards to Improve Teachers' Judgments—The Effect Shown in Ability to Score English Compositions" (*Journal of Educational Research*, Dec. 1925).

speech or essay must contain certain elements; but the relative importance of each has not been determined. Thus, it is agreed that a good composition must conform to certain standards of mechanical skill, as in matters of spelling, capitalization, punctuation, grammar; to standards of structural form, such as sentence structure, paragraph organization, facility of diction; to standards of thought content, whereby the writer through mechanics and form gives expression to a coherent train of thought.

Types of language tests.—Current tests in the field of language attempt to define and objectify the standards to be employed in judging the merit of an individual's language ability. In addition, the tests are more or less diagnostic, inasmuch as they measure the several aspects of the ability. The tests are of three general types: (1) those which measure the mechanics, such as grammatical correctness, punctuation, technical knowledge of grammar, sentence structure; (2) those which test ability to think of the proper word in a given situation; and (3) those for measuring general merit of the written composition. In this connection it is clear, of course, that the vocabulary tests described in the preceding chapter have a definite contribution to make, as well as the previously described tests of spelling.

Punctuation Scales

Several scales have been designed to test punctuation ability. The first to have wide use was the Starch Punctuation Scale.

STARCH PUNCTUATION SCALE

Description of the scale.—These scales are made up of a series of exercises arranged in "steps," each exercise consisting of sentences to be punctuated by the pupils. The sentences of each "step" of the scale are so selected that the difference in difficulty between any two successive steps of the scale is equal to the difference between any other two successive steps. Sample "steps" are as follows:

Step 7

1. I told him but he would not listen.
2. Concerning the election there is one fact of much importance.
3. The guests having departed we closed the door.
4. The train moved swiftly but Turner arrived too late.

Step 11

1. Paris Illinois is a smaller city than Paris France.
2. He asked what is the matter.
3. I like to work he said especially in the morning.
4. Chicago Illinois is a large city.

Method of using the scale.—The pupil is given a printed copy of the scale and instructed to correct the sentences by inserting the appropriate marks of punctuation where needed. The procedure is decidedly easy.

Function of the scale.—The scale measures the pupil's ability to use the correct form, the assumption being that if he can do this accurately in the scale, he will be able to apply the same knowledge in his own written work. On the other hand, his failures in the

scale can be easily classified and used as the basis for intelligent drill. Since the application of the scale is not at all difficult, and since the scoring is easy and therefore likely to be accurate, the use for class information is made simple. The value for diagnosis of this sort of mechanical difficulty for both elementary and high school pupils is evident. There is, however, one defect which limits the value of this test: the sentences used properly do not bring out in sufficient variety the many types of punctuation. However, the more important types are dealt with.

English Grammar Scales

STARCH'S ENGLISH GRAMMAR SCALES

Description of the scale.—The Starch Scales consist of three distinct series of exercises, constructed in the same general way as his punctuation scales, each set being arranged in a series of "steps." Each step consists of a group of exercises, requiring knowledge of definite language forms. The purpose is to determine whether the pupil can use the forms correctly rather than to test his knowledge of the rules on which they are based. Typical exercises are as follows:

Step 8. (Scale B)

1. The fact that I had never before studied at home, (I was at a loss; made me feel at a loss as to) what to do with vacant periods.
2. Both are going,—(he and she; him and her).
3. I don't believe I (will; shall) be able to go.

4. It is (the handsomest vase I almost ; almost the handsomest vase I) ever saw.

Step 11

1. There we landed, and having eaten our lunch (the steam-boat departed ; we saw the steamboat depart).
2. (After pointing ; when he had pointed) out my errors, I was dismissed.
3. The question of (whom ; who) should be leader arose.
4. He spoke to some of us,—namely (she and I ; her and me).

Method of using the scale.—Printed copies of the scale are given to the students, and they are instructed: “Each of the following sentences gives in parentheses two ways in which it may be stated. Cross out the one you think is incorrect or bad. If you think both are incorrect, cross both out. If you think both are correct, underline both.” The score is based on the highest step in which the pupil does seventy-five per cent of the exercises. Starch suggests the following scores as standards of attainment:

<i>Grade</i>	7	8	9	10	11	12	Freshmen
Score	8.0	8.3	8.6	8.9	9.2	9.5	10.3

Function of the scale.—As has been said, the scale is designed to test ability to use grammatical forms, not to test knowledge of rules of grammar or terminology of the subject. It carries out the idea that if the pupil has a sufficient knowledge of usage, the teacher need not attempt formal instruction in grammar. A large number of grammatical forms are included in the scale, and the instructor is therefore given ample opportunity to determine weaknesses of the pupils. The

scale itself has been criticized because it is not strictly diagnostic, in that the various forms are not arranged in a systematic manner, so as to make easy the recognition of types of weakness. However, the teacher will not find it difficult to make her own diagnosis for each pupil.

CHARTERS' DIAGNOSTIC LANGUAGE TESTS

Description of the scale.—W. W. Charters has designed scales to test the pupil's use of verbs, pronouns, and miscellaneous forms in their context, thus involving knowledge of good language usage. There are two forms of four parts each, one part being devoted to verbs, another to pronouns, and two to miscellaneous forms. Sample sentences are:

Them are my chums.
I and my sister went home.
Each of us were going.
I could of gone.
They dress so queer.

Method of using the test.—Printed blanks are given to the pupils. They are then instructed as follows: "This test is given to pupils who have studied language lessons to see how well they are able to tell when sentences are right and when they are wrong. Now look at the sample below:

I told him to go.

.....

"The plan is to read this sentence over carefully and see if it is right. If it is right, make a cross on the dotted line below the sentence. The sentence 'I told him to go'

is right, so we shall make a cross on the dotted line below it. Make the cross now.

"If the sentence is not right, we are to put the correct word or words on the dotted line below it. Let us try one that is not right, etc." After this careful instruction, the pupils proceed to work out the paper, being given ample time to finish, as the element of speed does not enter. A scoring key accompanies each form of the test, and careful directions are given for recording scores properly. Each scoring key is accompanied also by standards suggested by the author.

Function of the test.—The scale as described measures principally the ability to use correct forms of pronouns, verbs, and adverbs. Charters has also worked out one form in which the pupil is given an opportunity to give reasons for corrections; this tests the knowledge of the rule or principle governing the proper use of the form. As is indicated by the name, the teacher is expected to diagnose the pupil's difficulty and to undertake the proper instruction and drill. The author bases the tests upon his researches in language errors. Thus, he collected twenty-five thousand errors made by pupils in using pronouns in oral speech. Upon study, it developed that there were only forty different kinds of errors involved in the entire twenty-five thousand. The test is designed to bring in as many of these forty errors as possible and so to enable the teacher to determine those most likely to occur in her class in order to determine the direction of her drill. The scientific nature of the underlying basis of the test makes the resulting diagnosis a most useful one in laying the proper foundation work for usage in the matter of pronouns,

verbs, and adverbs. Incidentally, the test also brings out a number of miscellaneous weaknesses.

NEW YORK ENGLISH SURVEY TESTS

Description of the tests.—Although these tests go beyond the examination of bare grammar, they are included in the present category because they are intended to measure language usage in addition to technical knowledge of forms. The series includes a test of grammar, with norms for grades 7, 8, and 9; a test of language usage, having norms for grades 4 to 9; another test of sentence structure, provided with norms for grades 4 to 8; and a fourth part to test literature information, provided with norms for grades 7 to 10. The test of grammar is technical in nature, requiring the pupil to designate various parts of speech in given sentences and to identify types of sentences. The test of language usage has two parts: the first to examine the pupil's ability to choose correct expressions to replace incorrect forms in given sentences, and the second to test the pupil's ability to write into sentences the correct word which has been omitted, the omitted word being one that is frequently misused. The measurement of sentence structure is rather involved, inasmuch as the pupil is scored for essential structure, spelling, technical correctness and grammar, and language usage as demonstrated in his sentences written in answer to given questions. The fourth part, literature information, is composed of forty-eight multiple-choice exercises designed to "determine the pupils' general knowledge of the events and characters de-

scribed in and of the authors of the literature generally read by the pupils."

Method of using the tests.—Each section of the tests is contained in a separate booklet so that it is possible to administer one at a time. It is, of course, not necessary to make use of all four sections, since it may happen that a teacher will wish to examine her class in only one field. For a more nearly complete survey, however, the several sections are intended to be employed. Specific directions for administering, which are very simple, and time limits for each section are stated on each blank.

14. We never was to any large city before.
(were in) (was at) (went to) (were to)
15. The pictures were about them cities.
(these) (them there) (that) (this)

Language Usage

5. An author who wrote many poems for young children is
 - Rudyard Kipling
 - Robert Louis Stevenson
 - Sir Walter Scott
 - Philip Nolan
9. Alfred Tennyson was
 - an American
 - a Scotchman
 - a Canadian
 - an Englishman

Literature Information

Function of the tests.—With the exception of the section covering literature information, the tests examine the mechanical aspects of the use of English. No part of the test evaluates quality of expression. That task is left for the composition scales, and, no doubt, rightly so. The New York Survey Tests should be of service in the analysis of pupil difficulties in so

far as the proper use of common forms and the mastery of technical grammar are concerned. The value of the literature information test depends, of course, upon the merit of its contents from the point of view of universality of material. One section, sentence structure, is more than its name implies, for the pupil's sentences are graded for both form (mechanics) and comprehension. The pupil is required to write a specified number of sentences in response to a short statement or problem. Inasmuch as part of his score is dependent upon his comprehension of the situation, the measure becomes, in a sense, an "intelligence test." As previously stated, the introduction of extraneous factors is likely to confuse the results of a test.

CROSS ENGLISH TEST

Description of the test.—This test, intended primarily for high school seniors and college freshmen, has three similar and equivalent forms, each in eight parts as follows: spelling, pronunciation, recognizing a correct sentence, punctuation, verb forms, pronoun forms, idiomatic expressions, and miscellaneous faulty expressions. "The author has drawn upon his extensive observation of what young people say and write and has selected as far as possible" the "key" errors. The norms are based on results from 3,000 high school students and 3,500 college freshmen.

Method of using the test.—The entire test is contained in a single booklet. The time allowed is forty-five minutes. Inasmuch as instructions are printed for each part of the test, the student is told to begin with the

first, to do as many as he is sure of, and then to go over the remaining items in the time that is left.

Function of the test.—The test is intended to measure the individual's ability to employ "correct grammatical forms of English and acceptable sentence structure in speaking and writing." The unit in this test is the sentence, since it is regarded by the author as the fundamental unit. Like most other instruments in the field, it is designed for the analysis of difficulties, as well as for the determination of achievement levels.

Tests of Diction

TRABUE COMPLETION LANGUAGE SCALES

Description of the scales.—M. R. Trabue has worked out, on a scientific basis, a considerable number of scales, lettered from A to M, and also a Scale Alpha and a Scale Beta. In the main, they follow one general form. A sentence is given in which one or more words are left blank, and the pupil is required to fill in the missing word, thus completing the sentence. Scales B, C, D, E, and F are of equal value, and so may be used interchangeably to avoid danger of "coaching." Sample sentences from Scale D are:

1. We are going school.
5. Hard makes tired.
8. The best advice usually obtained one's parents.
10. a rule associations friends.

Method of using the scale.—Printed blanks are distributed to the pupils, and they are instructed to write one word in each blank, in each case writing that word which makes the best sense. Sample sentences are given for practice by the children so that the method of procedure may be entirely clear. A time limit is imposed for each scale, which must be carefully observed. A scoring key is furnished with the scales, which makes the determination of right and wrong usage very clear. Allowance is made for slight imperfections, so that half credit is allowed when the sentence is less than perfect, but still makes reasonably good sense.

Function of the scale.—Trabue describes the scale as “an attempt to derive one or more scales for the measurement of ability along certain lines closely related to language.” Other psychologists have described it as a test of “general ability” or of “intelligence.” Certainly the scale involves a knowledge of grammar, an ability to reason, to determine the fitness of various words in their relations, and a knowledge of diction and some range of vocabulary. Norms have been worked out for all grades of the grammar school and the four years of the high school. The scale is described as “for children between seven and twenty years old.” On account of its high correlation with tests of “general intelligence,” the scale may serve a double purpose, either in a rough way to indicate the intellectual level of the individual or the class, or to determine language development. As a test of language it may be used as prognostic, or as diagnostic, to determine corrective measures for the pupil or the

class; or it may be used as an achievement test in language, to find out the relative development of the language ability of the individual or group in question.

Tests of Composition

Up to this point tests and scales have been considered which are to be placed in the hands of the pupils themselves in order to discover their ability along lines involved in language usage; but these tests do not measure the completed composition. In addition, a number of scales have been devised which give standards for judging completed paragraphs or entire compositions on a basis of general quality, or of elements comprehended in the expression "general quality." These scales are made up of selected paragraphs indicating various grades of attainment in composition. The teacher is expected to compare the actual work of the pupil with the selections of the scale and to assign a value to the composition equal to that of the scale paragraph which it most nearly approaches in quality. Such scoring involves questions of judgment, and for this reason composition scales are not nearly so objective as are those hitherto described in this chapter. Accordingly, teachers have been divided in their opinions as to the value of composition scales. In general it may be said that those teachers who have practised the use of the composition scales have found them a genuine help in setting reasonable standards and thereby have been enabled to arrive at judgments of pupils' work much more accurate than

those reached by purely subjective estimates. The first use of such a scale is likely to be relatively inaccurate; but practice makes it a valuable adjunct to the teacher in both elementary and high schools.

NASSAU COUNTY SUPPLEMENT TO THE HILLEGAS SCALE

Milo B. Hillegas was the pioneer in devising a scale for measuring English composition, and his scale, published in 1912, has been made the basis for several succeeding scales. The general method of constructing the scale was to select a series of compositions, most of them actual work of school children, arranged in order of merit, as determined by consensus of a large number of competent judges and corrected by careful statistical methods. The original scale was not entirely satisfactory since the selections were of differing length and character, no two dealing with the same theme, and separated by irregular intervals. Several attempts have been made to correct these difficulties; Prof. E. L. Thorndike made an extension of the scale, and Prof. M. R. Trabue devised the Nassau County Supplement described below.

Description of the Nassau County Supplement to the Hillegas Scale.—The scale is intended for use in grades four to twelve inclusive. It consists of ten samples of composition ranging in value from 0 to 9.0, arranged in ascending order on one sheet. The actual values are: 0, 1.1, 1.9, 2.8, 3.8, 5.0, 6.0, 7.2, 8.0, 9.0. Thus they approximate closely a ten unit division of the scale. The first seven samples are compositions writ-

ten on the subject "What I Should Like to Do Next Saturday," and were obtained in a survey of the schools of Nassau County, N. Y. Specimen 10 is taken from literature. The entire scale is intended to measure only "general quality," there being no indication of the specific factors involved in this determination.

Method of using the scale.—The author gives directions for obtaining compositions for comparisons, suggesting that standard conditions be observed. Thus, the topic assigned "must be interesting and suggestive to the pupils, and at least twenty minutes must be allowed for the writing. 'What I Should Like to Do Next Saturday' will produce a higher average quality of results than 'How to Play Baseball,' but it will probably produce a lower average than 'The Most Exciting Experience of My Life.' "

The teacher is then told to compare the general quality of the composition with the general qualities of the various samples on the scale and to assign to the composition the numerical value of the printed sample which it most nearly equals in general merit.

Function of the scale.—As has been indicated, the purpose of the scale is to measure "general quality" of English composition. As there is no attempt to analyze general quality, it is valuable rather as a measure of general attainment of the pupil or of the class than as a diagnostic or prognostic measure. Teachers untrained in the use of the scale differ widely in results upon the first application of it, but experiments have demonstrated that after a few hours of training the judgments become more objective, and

variation between teachers becomes relatively slight. As a general measure it has definite value.

OTHER COMPOSITION SCALES

Other scales arranged on the same general plan as the Nassau and Hillegas Scales are the Breed and Frostic, especially adapted to the sixth grade; the Willing, for grades four to eight, which also involves the factors of story value, spelling, punctuation, capitalization, and grammar; and the Hudelson, which is a supplement to the Hillegas Scale based upon one thousand compositions written by first year high school pupils in Virginia. The Hudelson Scale gives a more uniform value from step to step than does the Nassau Supplement and is accompanied by complete directions for scoring compositions, which make a valuable addition to the language teacher's equipment.

A somewhat different type of measure is represented by the Harvard-Newton Scale, devised under the direction of F. W. Ballou. It is arranged in four parts, one for each of the forms: narration, description, exposition, and argumentation. Each of these parts forms a distinct scale, and so makes a valuable instrument where the composition work of a school is taught in accordance with these divisions. The idea involved in the Harvard-Newton Scale is also used in the Minnesota Composition Scale, devised by M. J. Van Wagenen. This scale is arranged in three parts, a separate scale for each of the forms: narration, de-

scription, and exposition. These scales are made up of either fourteen or fifteen selections, actual compositions written by Minnesota school children. Each contains compositions written upon the same general topic, that in narration being "When Mother Was Away," in description, "It Was a Sight Worth Seeing When the Troops Marched By," and for exposition, "How I Earned Some Money." Each sample is given a definite value on each of three factors: "structure," "mechanics," and "thought content." In this way it becomes possible to compare and evaluate work done along three distinct lines, thus permitting diagnosis to some extent, a thing not achieved in all composition scales.

The Seaton-Pressey Minimal Essentials Scale and the Pressey Diagnostic Tests in English Composition are frequently mentioned in connection with tests of composition. It seems, however, that they belong more appropriately in the same group with the New York Survey Tests, the Cross English Test, and others which measure the mechanical aspects of composition, such as capitalization, punctuation, grammar, and sentence structure. The Seaton-Pressey and the Pressey scales do not measure the quality of the finished composition.²

Mention should also be made of the Lewis English Composition Scales which are devised to measure ability in the writing of business and social correspondence, including the following: order letters, letters of

² See "English Composition: Report of the Fourth Annual Nation-wide Testing Survey" (Bloomington, Ill., The Public School Publishing Company).

application, narrative social letters, expository social letters, and simple narration.

Ability to Judge Poetry

ABBOTT AND TRABUE EXERCISES IN JUDGING POETRY

Description of the test.—If the art of poetry is to be included in the curriculum for the purpose, among others perhaps, of training critical appreciation, it would be desirable, according to many, to have “an objective test of independent critical judgment applied to specimens of the art previously unknown.” If such a measuring scale were practicable, the success of the teaching of poetry might be revealed. This is the purpose for which the Abbott and Trabue test is designed. From more than one hundred sets of poems, each in four versions, the authors submitted two graduated series, X and Y, of thirteen sets each, and of approximately equal difficulty, to 3,500 judges, including persons of all grades, from the university to the fifth grade. The original poem, in each case, was “deliberately made worse by diminishing or completely annulling one or another of its characteristic merits,” such as the rhythmic form, imaginative or emotional quality. The poems retained in the final form are those each of which has been designated as superior to all its variants by a group of experts, such as poets, literary editors, critics, and professors of literature. The range of poetic types is wide, from Mother Goose to Browning, from Milton to Amy Lowell.

Method of using the tests.—Each pupil is given a

booklet containing the 13 sets of poems. The authors advise using both sets, requiring forty minutes each. "Read the poems A, B, C, D, trying to think how they would sound if read aloud. Write 'Best' on the dotted line above the one you like best as poetry. Write 'Worst' above the one you like least."

Set 4. *Tea Time*

A (.....)

The coals beneath the kettle croon
And clap their hands and dance in glee;
And even the kettle hums a tune
To tell you when it's time for tea.

B (.....)

Sweet coal, I hear thy tender croon,
I see thee dancing in thy glee;
Dear kettle, I thank thee for the tune
That bids my heart prepare for tea.

C (.....)

The coal when it is lighted glows
Enough for anyone to see;
By watching the kettle, everyone knows
When it is time to come to tea.

D (.....)

The coal that's under the kettle does crackle
And dances up and down in its glee;
And the teakettle sings its gay tune
That says that it's almost time for tea.

Function of the tests.—The purpose of these tests is to determine the appeal that various types of poetry make to individuals of various ages and grades. The tests avoid such issues as the nature of poetry itself and the relative merits of the several "schools" of poetry.

Statistical treatment has shown these tests to be of little or no value as *measuring* devices in the elementary school, for they begin to be significant only in the upper part of the high school and in college. It is possible, however, that though the tests are of little value to most teachers for purposes of measurement, they may nevertheless serve a useful purpose as teaching devices.

Conclusions.—When the entire field is surveyed there is every ground for believing that the teacher who has familiarized herself with the various types of measuring instruments described in this chapter will be much better equipped as a teacher of the English language than if she had not gone into the study of structure, mechanics, and content involved in the use of these scales.

Yet it should be clear that, with the exception of certain of the composition scales, the measures in the field of language are devoted very largely to the formal aspects of the subject. The tests in this subject suffer from certain limitations: (1) they do not measure adequately the active use of language, that is, the functional aspect of language usage; (2) they can at best be only samplings of language and grammar; and (3) some do not possess as high a degree of consistency (the statistical term is "reliability") as may be de-

manded. In spite of these defects, however, there can be no question that the tests and scales have made a distinct contribution to the more effective teaching of language and the more precise measurement of the product.

MATERIALS NEEDED

Abbott and Trabue, Exercises in Judging Poetry. (Bureau of Publications, Teachers College, Columbia University.) Single copy 5 cents; manual of directions 25 cents.

Ballou, F. W., Harvard-Newton Composition Scale. (Harvard University Press, Cambridge, Massachusetts)

Charters, W. W., Diagnostic Language Test, for grades 3 to 8. (Bloomington, Ill., The Public School Publishing Company.) Sample set, 10 cents; 80 cents per hundred.

Cross English Test, (Yonkers-on-Hudson, N. Y., World Book Company). \$1.20 per 25.

Hudelson Typical Composition Ability Scale, (Bloomington, Ill., The Public School Publishing Company). \$1.00 per 25.

Lewis, E. E., Scales for Measuring Special Types of English Composition, for grades 3 to 12; (Yonkers-on-Hudson, N. Y., World Book Company). Booklet containing all five scales, 25 cents.

New York English Survey Tests. (Bloomington, Ill., The Public School Publishing Company.) \$1.00 per 100.

Seaton-Pressey Minimal Essentials Scale, 75 cents per 100. Pressey Diagnostic Tests in English Composition; (a) and (b) 75 cents per 100; (c) and (d) \$1.50 per 100. (Bloomington, Ill., The Public School Publishing Company.)

Starch, D., Punctuation Scale (Daniel Starch, 1374 Massachusetts Ave., Cambridge, Mass.) 80 cents per 100.

Starch, D., English Grammar Scale. (Daniel Starch, 1374 Massachusetts Ave., Cambridge, Mass.)

Trabue, M. R., Completion Test Language Scales, for grades 2 to 12 and above. Eleven different forms, B, C, D, E, and F are equivalent, L and M are equivalent and especially intended for use in high school. J and K are equivalent and for

use with college students and adults. Alpha and Beta are equivalent and are combinations of the other tests. A Key to the Trabue Language Scales (A Manual) 75 cents. Key for forms B and C 20 cents. Sample copy of each scale 10 cents. Scales Alpha and Beta \$1.25 per hundred, all others 50 cents per hundred. (New York, Bureau of Publications, Teachers College, Columbia University.)

Trabue, M. R., Nassau County Supplement (New York Teachers College, Columbia University). Single copy 8 cents; booklet, 35 cents.

Van Wagenen, M. J., Minnesota Composition Scale, (Minneapolis, M. J. Van Wagenen, University of Minnesota). Booklet complete, 25 cents.

Willing Scale for Measuring Written Composition, (Bloomington, Ill., The Public School Publishing Company). Single copy 9 cents; three or more 6 cents.

SUPPLEMENTARY LIST OF TESTS

Barrett-Ryan Literature Test.³

High school grades.

Briggs English Form Test.⁴

7th grade and above.

Clapp's Test for Correct English.⁵

Grades 5 to 12.

Clapp-Young English Test.⁵

Grades 5 to 12.

Clark Letter-Writing Test.⁶

Grades 5 to 12.

Columbia Research Bureau English Test.⁷

Upper high school grades and college.

Denver Curriculum Tests in English.⁸

Grades 7, 8 and 9.

³ Bureau of Educational Measurements and Standards, Teachers College, Emporia, Kansas.

⁴ Bureau of Publications, Teachers College, Columbia University.

⁵ Houghton Mifflin Company, Boston, Mass.

⁶ The Public School Publishing Company, Bloomington, Ill.

⁷ World Book Co., Yonkers, N. Y.

⁸ Denver Public Schools, 414 Fourteenth St., Denver.

Franseen Diagnostic Test in Language.⁹

Grades 3 to 8.

Kennon Test of Literary Vocabulary.⁴

Grades 11 and 12.

Kirby (Iowa) Grammar Test.¹⁰

Logasa-McCoy-Wright Tests for Appreciation of Literature.⁶

Grades 11 and 12.

Markham English Vocabulary Test for High School and College Students.⁶

Schutte English Diction Test.⁶

Principally for high school grades.

Tressler English Minimum Essential Test.⁸

Van Wagener Reading Scales in English Literature.⁶

Grades 8 to 12.

Wakefield Diagnostic English Test.⁹

Wilson Language Error Test.⁷

SELECTED REFERENCES

Ashbaugh, E. J., "The Measurement of Language" (*Journal of Educational Research*, June, 1921).

Ashbaugh, E. J., "Senior High School English as Revealed by a Standardized Test" (*Journal of Educational Research*, Vol. 13, April, 1926, pp. 249-258).

Brown, M. D., and Haggerty, M. E., "The Measurement of Improvement in English Composition" (*The English Journal*, Vol. 6, pp. 515-527, 1917).

Fillers, H. D., "Oral and Written Errors in Grammar" (*Educational Review*, Vol. 54, pp. 458-470, 1917).

Hillegas, M. B., "Hillegas Scale for Measurement of English Composition" (*Teachers College Record*, September, 1912).

Hosic, J. F., "The Essentials of Composition and Grammar" (*Fourteenth Yearbook*, National Society for the Study of Education, 1915).

Hosic, J. F., "Chicago Standards in Oral Composition for

⁹ College of Education, University of Cincinnati.

¹⁰ Bureau of Educational Research and Service, University of Iowa, Iowa City.

Grades 6 to 8" (*Elementary English Review*, Vol. 2, No. 5, p. 107 and No. 6, p. 255, 1925).

Hudelson, E., "English Composition, Its Aims, Methods, and Measurement" (*Twenty-Second Yearbook*, National Society for the Study of Education, pp. 75-82, 1923).

Hudelson, E., "Use of Objective Standards to Improve Teachers' Judgments—The Effect Shown in Ability to Score English Compositions" (*Journal of Educational Research*, December, 1925).

Inglis, A., *Principles of Secondary Education*, Chapter 12 (Boston, Houghton Mifflin Company, 1918.)

Irmina, Sister M., "A Study of Language and Grammar Tests," *Catholic University, Educational Research Bulletin*, Vol. 1, no. 8 (Washington, D. C., Catholic Education Press, 1926.)

Klapper, P., "Some Observations Concerning the Measuring of Ability in Composition." *Contributions to Education*, Vol. I, Chapter 6 (Yonkers-on-Hudson, N. Y., World Book Company).

Starch, D., "The Measurement of Ability in English Grammar" (*Journal of Educational Psychology*, Vol. 6, 1915).

Trabue, M. R., "The Nassau County Supplement to the Hillegas Scale" (*Teachers College Record*, January, 1917).

Twenty-Second Yearbook of the National Society for the Study of Education; "English Composition, Its Aims, Methods, and Measurements" (Bloomington, Ill., The Public School Publishing Company).

Willing, M. H., "Individual Diagnosis in Written Composition" (*Journal of Educational Research*, Vol. 13, No. 2 February, 1926, pp. 77-89).

CHAPTER IX

ARITHMETIC

Importance and problems of instruction in arithmetic.

—In importance, arithmetic is probably second only to reading. Mastery of the four fundamental processes is one of the essentials for the satisfactory conduct of affairs in school and out. While it may be true that emphasis has at times been wrongly placed in the teaching of arithmetic, the fact remains, nevertheless, that it justly occupies a central position in the curriculum of the school, particularly during the earlier years. What has been said applies to the fundamentals of arithmetic as taught in the primary and elementary grades; it does not necessarily hold for high school or college mathematics, which present another problem that does not warrant consideration at this time.

Because of its importance, the subject of arithmetic has been investigated and experimented upon more than any other branch of instruction, with the possible exception of reading. Yet, formerly arithmetic had been regarded as a rather simple subject with respect to content and methods of instruction. But experimental investigation has demonstrated that it is a highly complicated subject, giving rise, at the same time, to such problems as: (1) the nature and elements of arithmetical ability; (2) the content of arithmetical instruction and the age levels at which certain proc-

esses should be taught; (3) methods of teaching; (4) methods of measuring arithmetical ability; and (5) utilizing the results of measurements.

By means of careful experimentation, the important topics of instruction in arithmetic may be determined.¹ Once determined, however, there yet remains the difficult task of deciding at what grade the various materials shall be taught so as best to suit the pupils' level of mental development. It is in this latter problem that standardized tests of arithmetic ability may be of considerable value, for there is a wide diversity of practice from school system to school system. More results, however, have been achieved in the first problem, that of discovering the significant elements of instruction. For example, experiments have been conducted to find the relative difficulty of various number combinations in the four fundamental processes,² to determine types and frequencies of errors of mechanics,³ to determine the optimum amount of practice for different topics,⁴ to discover the relationship between the various aspects of arithmetic,⁵ etc.

Types of arithmetic tests.—Arithmetical computation involves, in broad outline, number concepts;

¹ See Thorndike, E. L., *The Psychology of Arithmetic*, pp. 75-95. Also "The Psychology of Drill in Arithmetic: The Amount of Practice" (*Journal of Educational Psychology*, Vol. 13, No. 4, pp. 183-194, 1921).

² Clapp, F. L., "The Number Combinations, their Relative Difficulty and the Frequency of their Appearance in Textbooks," *Bulletins Nos. 1 and 2, Bureau of Educational Research*, University of Wisconsin, 1924.

³ Buswell, Guy T., and John, L., "Diagnostic Studies in Arithmetic," *Supplementary Educational Monographs*, No. 30; University of Chicago, 1926.

⁴ Thorndike, E. L., "The Psychology of Drill in Arithmetic; The Amount of Practice."

⁵ Buswell, Guy T., and Judd, C. H., "Summary of Investigation relating to Arithmetic," *Supplementary Educational Monographs*, No. 27; University of Chicago, 1925.

ability to read, write, and speak the number symbols; associations established between number combinations (the fundamental processes); reasoning and number association, this last being what is ordinarily known as arithmetical reasoning. Clearly, then, the measurement of efficiency in arithmetic consists principally in measuring the accuracy and rate with which the arithmetic processes and applications are employed at the various levels of difficulty. In test construction, therefore, a distinction has been made between measuring the fundamental operations alone and problem solving. The justification of separating tests of fundamental processes from those of arithmetic reasoning in the solution of problems has been well demonstrated.⁶ There are, consequently, measures of (1) fundamental operations, and of (2) arithmetical reasoning. Of the former type there are, among others, the Courtis Standard Research Tests and the Cleveland Survey Tests; of the latter there are, for example, the Monroe Reasoning Tests and the Stone Reasoning Test.

COURTIS STANDARD RESEARCH TESTS—SERIES B

Description of the tests.—These tests consist of a series of problems in each of the four fundamental operations. There are twenty-four problems of equal difficulty in each operation. In Addition the class is allowed eight minutes to solve as many problems as possible; four minutes are allowed for the problems in

⁶ Stone, C. W., "Arithmetical Abilities and Some Factors Determining them." *Contributions to Education*, No. 10, Teachers College, Columbia University, 1908. Also Buswell and Judd, *op. cit.*

**SAMPLE PROBLEMS FROM THE COURTIS STANDARD
RESEARCH TESTS**

Eight of the Twenty-Four Problems in Addition

You will be given eight minutes to find the answers to as many of these addition examples as possible. Write the answers on this paper directly underneath the examples. You are not expected to be able to do them all. You will be marked for both speed and accuracy, but it is more important to have your answers right than to try a great many examples.

927	297	136	486	384	176	277	837
379	925	340	765	477	783	445	882
756	473	988	524	881	697	682	959
837	983	386	140	266	200	594	603
924	315	353	812	679	366	481	118
110	661	904	466	241	851	778	781
854	794	547	355	796	535	849	756
965	177	192	834	850	323	157	222
344	124	439	567	733	229	953	525

Eight of the Twenty-Four Problems in Subtraction

92971900	104339409	60472960	119811864
62207032	74835938	50196521	34379846
137769153	144694835	123822790	80836465
70176835	74199225	40568814	49178036

Ten of the Twenty-Five Problems in Multiplication

6385	8736	5942	6837	4952
48	502	39	680	47
3876	9245	7368	2594	6495
93	86	74	25	19

Eight of the Twenty-Four Problems in Division

25) 6775	94) 85352	37) 9990	86) 80066
73) 58765	49) 31409	68) 43520	52) 44252

Subtraction, six for the problems in Multiplication, and eight for Division. The problems of each of the four operations are arranged on separate pages of the test folder.

Scoring the tests.—Score cards are provided to facilitate the marking of the papers. Only those problems with correct answers receive credit, but the number attempted is also recorded as a basis for determining accuracy. The scores are recorded on a class record sheet which is furnished with the test. Complete directions are supplied for computing median scores and median differences. The method is somewhat complicated but carefully described so that the class-room teacher may follow the directions step by step. Tentative standards are given as follows for June.

STANDARDS FOR COURTIS STANDARD RESEARCH TESTS

GRADE	Test 1 ADDITION	Test 2 SUBTRACTION	Test 3 MULTIPLICATION	Test 4 DIVISION
3	4	5	0	0
4	6	7	6	4
5	8	9	8	6
6	10	11	9	8
7	11	12	10	10
8	12	13	11	11

Functions of the tests.—The Courtis Tests are among the best known of the educational tests. Arithmetic lends itself to exact measurement; hence it was one of the first of the school subjects for which tests were constructed. The tests provide an easy means whereby the teacher may compare her class with what should be ex-

pected for that grade. They also provide a very practical teaching device for drill in the fundamental operations without disturbing the schedule of instruction for the other children in the grade.

THE COURTIS STANDARD PRACTICE TESTS IN ARITHMETIC

Description of the test.—These are not ordinary tests but a series of carefully graded problems for daily practice. The material for practice consists of two parts: (1) a set of lesson cards and (2) a student's record and practice pad. The lesson cards are a series of problems in addition, subtraction, multiplication, and division. The first card, for example, contains 72 simple problems in addition. The answers to the problems are printed on the back of the card. The practice pad contains sheets of transparent paper, and one of the lesson cards is inserted under a sheet of paper so that the pupil may work the problem on the transparent paper.

On the first day a preliminary test is given, and all children reaching a certain standard are excused from drill. Those pupils who need the drill spend from three to four minutes each day in practice on each lesson card; the amount of time spent in practice depends upon the school grade. The pupil is trained to score his own work by comparing his answers with those given on the back of the cards. The score made each day is recorded, and graphs are drawn by the child so that the pupil may see the effect of his own practice. Practice is continued on any card until standard efficiency is

**SAMPLE LESSON OF THE COURTIS STANDARD
PRACTICE TESTS**

Lesson No..... Form..... Date.....

Name..... Grade.....

21	81	31	71	51
14	12	16	13	17
—	—	—	—	—

61	41	51	71	31
15	18	19	52	27
—	—	—	—	—

41	51	61	71	81
24	45	26	69	23
—	—	—	—	—

10
B

51	22	32	42	52
38	22	31	23	34
—	—	—	—	—

,

62	72	82	52	73
32	33	34	44	23
—	—	—	—	—

SCORES	COURTIS STANDARD PRACTICE TESTS	TRIAL No.
Tried —		
Right —	LESSON No. 10 MULTIPLICATION	Form B

reached. The pupil then passes on to the next practice card. This necessitates individual instruction in arithmetic, but the practice cards make this possible. Each pupil may progress at his own rate. There are forty-eight cards containing lists of problems arranged by gradual steps for daily practice. The pupil can be drilled on these problems, and tests are provided from time to time for measuring progress.

The Teacher's Manual describes in detail methods for teaching each of the simple operations in arithmetic. There are two forms, A and B, of the Standard Practice Tests of equal difficulty and, therefore, interchangeable.

The practice tests have been criticized⁷ in that the arrangement of the number combinations provides an equal amount of drill on all the combinations, some of which are much more difficult than others for the child; therefore, more practice should be provided for the more difficult combinations and less for the easier ones. For examples, 8 and 9 need more practice than combinations of 1 and 2. This criticism seems valid, and no doubt the material of the tests should be revised on the basis of the difficulty of the number combinations.

COMPASS DIAGNOSTIC TESTS IN ARITHMETIC

Description of the tests.—This series of twenty tests is perhaps the most exhaustive of the measures in the

⁷ Osburn, W. J., "A Study of the Validity of the Courtis and Studebaker Tests in the Fundamentals of Arithmetic" (*Journal of Educational Research*, Vol. VIII, No. 2, September, 1923).

field. There are, for instance, separate tests of addition of whole numbers, of fractions, of mixed numbers; a similar group for each of the other three operations; tests of denominate numbers, of mensuration, interest, business forms, definitions, rules and vocabulary; and there are tests of problem analysis. The range is from grade 2 to grade 8. It is clear that the twenty tests cover all the phases of elementary school arithmetic.

The Compass Diagnostic Test may be given at any time of the year, especially with a view to measuring the results of instruction in a particular unit, or to gain adequate information about all or some members of a new incoming group. It is not expected that any one school will necessarily use all twenty tests. Rather, as the name suggests, the measure should be used for diagnostic purposes. Grade norms and age equivalents of total scores are provided.

Function of the tests.—It is the view of the authors of these tests that most other measures in the field of arithmetic are of the survey type, and too little of the diagnostic; they yield results showing only the general level of ability for a class or a pupil. It is the purpose of the Compass Tests, however, to isolate the specific skills and difficulties for measurement.

It is doubtful whether the authors' sweeping criticisms of all other tests is justified, for some of them possess distinct diagnostic qualities. However, there can be but little question that the Compass Tests are the most comprehensive of the group; and this greater comprehensiveness no doubt increases their value where a thorough and searching analysis is desired.

CLEVELAND SURVEY ARITHMETIC TESTS

Description of the tests.—These tests, devised by Judd and Counts for a survey of the Cleveland schools, consist of fifteen sets of exercises in the four fundamentals of arithmetic and fractions.

These sets are arranged in spiral form, that is, the same operations recur several times in the test, but each time the problems introduce greater complexity. The first problems in addition consist of a set of two single place numbers. These are followed by sets of simple problems in subtraction, multiplication, and division; then in turn by a set of problems consisting of five single place numbers to be added. This spiral arrangement is followed throughout the test, each new set of problems being more difficult than the preceding. The eighth and fifteenth sets are problems in the addition, subtraction, multiplication and division of fractions.

Scoring the test.—The pupils are given from thirty seconds to 4 minutes, differing for the various tests, in which time they are to work as many problems as they can in each set. The total actual working time for the fifteen sets of problems is twenty-two minutes. The problems are scored by means of a key. The number right in each set constitutes the final score. There are no general standards or norms for these tests, for only separate standards for St. Louis and Grand Rapids are given on the folder accompanying the tests. There is some variability between these scores, those of St. Louis being in general the higher. The nature of the norms must be remembered where the test is used.

STANDARDS (MEDIAN NUMBER OF EXAMPLES CORRECT) FOR THE CLEVELAND SURVEY ARITHMETIC TESTS

ST. LOUIS, MISSOURI

Grades

TEST	3-B	3-A	4-B	4-A	5-B	5-A	6-B	6-A	7-B	7-A	8-B	8-A
A	14.6	18.3	19.8	21.3	22.5	22.5	26.3	26.4	27.8	28.4	32.3	32.2
B	9.9	12.2	17.1	17.0	18.0	20.0	20.3	20.6	22.8	24.2	26.7	28.3
C	7.6	10.5	16.7	15.4	16.9	16.7	18.2	18.3	18.9	19.8	20.7	21.9
D	9.0	12.2	15.8	16.3	18.4	17.8	19.3	20.5	21.3	22.3	23.8	25.7
E	3.8	4.8	5.7	5.4	6.0	6.1	6.9	7.1	6.6	7.4	8.0	8.4
F	2.3	3.5	5.6	6.0	6.4	7.4	8.0	8.3	8.5	9.6	10.1	11.3
G	2.7	3.5	4.9	5.1	5.5	5.6	5.9	6.2	6.4	6.9	7.4	7.8
H	0.7	3.8	7.8	6.8	4.8	6.5	8.0	8.1	9.5	9.7	10.8	12.0
I	1.1	1.4	2.0	2.0	3.0	3.2	3.9	4.1	4.5	5.0	5.4	5.8
J	1.6	2.9	3.8	3.9	4.1	4.3	5.0	5.1	5.2	5.3	5.4	5.8
K	3.3	4.0	5.0	5.8	6.9	7.4	8.3	9.7	10.3	11.7
L	2.5	2.9	3.1	3.4	4.3	4.1	4.6	4.7	5.2	5.3
M	0.5	2.1	2.9	3.3	3.4	3.7	4.2	4.4	4.5	4.9	5.2	5.3
N	0.8	1.1	1.3	1.4	1.6	1.8	2.0	2.0	2.6	2.7
O	2.5	3.3	3.3	3.6	4.1	4.8	5.6	6.1	6.6

GRAND RAPIDS, MICHIGAN

Grades

TEST	3-B	3-A	4-B	4-A	5-B	5-A	6-B	6-A	7-B	7-A	8-B	8-A
A	11.8	13.4	13.6	16.4	20.3	21.5	22.8	25.0	26.5	27.3	29.5	30.3
B	6.3	8.4	9.1	12.1	14.7	15.9	16.8	19.1	21.3	20.7	22.8	25.5
C	7.1	11.3	13.7	14.0	15.5	17.0	17.7	18.8	19.3	20.7
D	6.9	10.4	12.5	14.3	15.5	16.9	18.4	19.7	20.5	23.0
E	4.1	4.6	5.2	5.4	6.0	6.6	7.2	7.2	7.8	8.1
F	2.8	4.1	6.0	6.5	7.1	8.0	9.3	9.6	10.3	11.0
G	2.2	3.3	4.5	4.9	5.3	5.6	6.1	6.1	6.7	6.8
H	6.3	6.2	6.5	9.0	7.8	8.6	8.8
I	0.7	0.9	1.3	1.4	2.3	3.0	3.8	4.1	4.0	4.7
J	2.8	3.4	3.7	4.1	4.5	5.4	5.3	5.7	6.5
K	3.0	4.3	5.4	6.5	7.5	8.8	9.7	10.3
L	2.3	2.9	3.3	3.6	4.3	4.5	4.9	4.9
M	2.3	3.0	3.6	4.3	4.5	4.9	5.0	5.7	5.7
N	0.7	0.8	1.1	1.4	1.7	1.8	2.0	2.3
O	3.5	3.6	3.9	4.6	5.5	4.8

Function of the tests.—The Cleveland Survey Tests are highly diagnostic, more so than the Courtis Tests, since they analyze the fundamental processes into successive steps. The teacher may determine what types

of problems the pupil can solve in addition, for example, and what types he can not solve. This gives an indication of the pupil's difficulties. Through the use of the tests the teacher may determine where wrong

SAMPLE PAGE FROM THE CLEVELAND SURVEY ARITHMETIC TESTS

SET L—Multiplication—				Ats.	Rts.
8246	3597	5739	2648		
29	73	85	46		
—	—	—	—		
4268	7593	6428	8563		
37	64	58	207		
—	—	—	—		
SET M—Addition—					
7493	8937	8625	2123	5142	3691
9016	6345	4091	1679	0376	4526
6487	2783	3844	5555	4955	7479
7591	4883	8697	6331	9314	2087
6166	1341	7314	6808	5507	8165
—	—	—	—	—	—
5226	9149	6268	9397	7337	8243
2883	8467	7725	6158	2674	6429
2584	0251	8331	3732	9669	9298
0058	7535	5493	4641	5114	7404
2398	5223	3918	7919	8154	2575
—	—	—	—	—	—
SET N—Division—					
67) 32763	48) 28464	97) 36084	59) 29382		
78) 69888	88) 34496	69) 40296	38) 26562		
—	—	—	—		

principles are involved, where drill should be placed, and where drill need not be emphasized. The increasing complexity of the problems makes the tests especially suitable for this purpose.

WOODY-McCALL MIXED FUNDAMENTALS

Description of the test.—This test consists of thirty-five problems in addition, subtraction, multiplication and division of whole numbers, common and decimal fractions. The different operations are arranged in irregular order on the test sheet. In some of the problems the operation is indicated by words and in others by the usual signs. The pupil is allowed twenty minutes in which to solve as many problems as possible.

Scoring the test.—The score is the number of problems having correct answers expressed in lowest terms. Norms are given for October scores as follows:

WOODY-McCALL STANDARDS

<i>School Grade</i>	3	4	5	6	7	8
Standard Score	6.8	13.1	17.8	22.5	25.9	27.8

In order to make these standards comparable with results obtained later in the school year for each month after October, the following increments should be made.

<i>School Grade</i>	3	4	5	6	7	8
Increment to be added for each mo.	.54	.43	.42	.24	.25	.20

Separate norms are also provided for high and low sections of each grade.

SHOWING TWENTY OF THE THIRTY-FIVE EXERCISES OF
THE WOODY-McCALL MIXED FUNDAMENTALS TEST

WOODY-McCALL MIXED FUNDAMENTALS FORM 1

Name..... Age.....

Grade.....

Get the right answer to as many examples as you can in 20 minutes. Do all work on the front or back of this sheet.

(1) ADD 2 3 —	(2) $2 \times 3 =$	(3) 3) 6	(4) SUBTRACT 2 1 —	(5) MULTIPLY 23 3 —	(6) SUBTRACT 13 8 —
---------------------------	-----------------------	--------------	--------------------------------	---------------------------------	---------------------------------

(7) ADD 17 2 —	(8) $3 + 1 =$	(9) SUBTRACT 16 9 —	(10) MULTIPLY 254 6 —	(11) $4 \div 2 =$
----------------------------	------------------	---------------------------------	-----------------------------------	----------------------

(12) ADD 23 25 16 —	(13) SUBTRACT 393 178 —	(14) 2) 13	(15) ADD 9 24 12 15 19 —	(16) MULTIPLY 5096 6 —
------------------------------------	-------------------------------------	----------------	---	------------------------------------

(17) $2 \frac{3}{4} - 1 =$	(18) ADD \$12.50 16.75 15.75 —	(19) MULTIPLY 7898 9 —	(20) $\frac{1}{4}$ of 128 =
-------------------------------	---	------------------------------------	--------------------------------

Function of the test.—In many respects this test is not different from the others in the fundamentals of arithmetic. The miscellaneous arrangement of the problems probably more nearly simulates ordinary test conditions. The use of signs in many of the problems emphasizes the importance of a knowledge of their meaning. The score is a total of all problems solved correctly. The test is, therefore, to be used principally as a general survey test. Only by further analytical study of the items can it be used for diagnostic purposes. This limits the usefulness of the test for classroom purposes.

The Woody Arithmetic Scales are similar to the Woody-McCall Mixed Fundamentals except that in the former the problems of each operation are printed on separate pages of a folder, and norms are given for each of the fundamental operations.

MONROE DIAGNOSTIC TESTS IN ARITHMETIC

Description of the tests.—These tests are so similar to the Cleveland Survey Tests that only the chief differences between them will be pointed out. Monroe has twenty-one sets of problems in his tests, some of which are more difficult than the Cleveland Survey Tests. They are printed in four separate parts. The fourth grade uses the first two parts; the fifth grade uses the first three; and the sixth, seventh, and eighth grades use all four. The time allowances and method of scoring are similar to the Cleveland Tests. Tentative mid-year norms for the number of examples correct are given as follows:

MONROE DIAGNOSTIC TESTS IN ARITHMETIC

*Tentative Standards—Mid-year Scores
Number of Examples Correct*

<i>Test</i>	<i>Grade</i>				
<i>No.</i>	IV	V	VI	VII	VIII
1	8.3	8.5	10.2	12.0	12.7
2	3.0	5.3	7.1	8.0	8.9
3	2.2	2.7	4.0	4.6	5.2
4	1.1	1.8	2.3	3.4	4.0
5	2.3	2.7	3.3	3.4	4.0
6	1.1	1.6	2.6	3.3	4.5
7	2.2	2.8	3.4	3.9	4.3
8	1.2	2.3	3.1	4.0	4.4
9	2.7	5.8	6.5	7.5	8.2
10	1.4	1.9	3.4	3.9	5.4
11	.9	1.1	1.6	2.0	2.3
12		1.4	3.5	4.3	5.4
13		1.6	2.5	3.3	3.7
14		1.9	3.8	5.2	5.1
15		1.4	2.7	3.3	3.3
16		1.9	3.4	5.7	6.1
17			1.6	2.2	2.5
18			8.3	9.5	11.0
19			2.4	3.4	3.5
20			8.5	10.0	11.0
21			1.7	2.2	2.4

The Monroe General Survey Scale in Arithmetic is also similar, except that the purpose of this scale is to provide a single score which will be a general measure of the pupils' ability to perform the operations of arithmetic, and, as its name indicates, it is for survey purposes, not diagnostic.

MONROE'S STANDARDIZED REASONING TESTS IN
ARITHMETIC

Description of the tests.—These tests consist of a series of practical problems such as are to be found in an ordinary textbook in arithmetic. They involve the fundamental operations, fractions, denominate numbers, and percentage. They are not time tests, for the pupil is given sufficient time to solve as many problems as he can. There are three separate tests, one for grades 4 and 5, another for grades 6 and 7 and another for grade 8.

SAMPLE TEST II FOR GRADES 6 AND 7. FORM 1

STANDARDIZED REASONING TEST IN ARITHMETIC

Devised by Walter S. Monroe

6. Four loads of hay are to be put into a barn. The first load weighs 1.125 tons; the second, 1.75 tons; the third, 1.8 tons; the fourth 1.9 tons. Find the weight of the four loads.

P = 1

C = 2

7. A baker used $\frac{3}{5}$ lb. of flour to a loaf of bread. How many loaves could he make from a barrel (196 lbs.) of flour?

P = 3

C = 2

Scoring the tests.—The problems are to be scored for both correct principle (P) and correct answer (C). For example, if a mistake is made in an addition, but the operation is the correct one, a score is given for the correct principle. A key is provided as a guide in scoring the papers. Norms for the tests are as follows:

Form 1. GRADE NORMS FOR MID-YEAR TESTING

	Test I		Test II		Test III
	Grades		Grades		Grade
	IV	V	VI	VII	VIII
Correct Principle					
25-percentile	6.2	12.1	10.0	13.8	11.5
Median	11.3	19.2	14.2	19.7	17.2
75-percentile	16.8	25.9	19.4	24.7	22.8
* Rate					
25-percentile	5.2	8.0	6.4	8.0	5.3
Median	7.8	11.2	8.7	11.2	7.5
75-percentile	8.1	15.1	12.1	14.5	10.9
Correct Answers					
25-percentile	4.1	7.1	6.9	9.4	5.1
Median	7.0	11.3	10.4	13.4	9.0
75-percentile	10.7	15.5	14.0	17.4	13.0

* Sum of correct principle values of problems done correctly within ten minutes.

Function of the tests.—These tests fill a very definite need for a standardized test in problem solving. The child may be able to perform the separate operations without knowing how to apply them to the solution of problems. Such application of the various operations is the practical measure of success in arithmetic. A larger element of reasoning is required in applying arithmetical operations to problems than in merely performing the operations themselves, for not only is specific arithmetic ability requisite, but *general intelligence* is required as well. Although theoretical difficulties may be involved in such combinations of factors, they may be disregarded in considering the practical value of tests such as these, for the aim of instruction in arithmetic is concerned with the solution

of problems, not merely with the mastery of the four operations.

NEW STONE REASONING TEST IN ARITHMETIC⁸

Description of the test.—This test, in two equivalent forms, consists of twenty-one problems of increasing difficulty and is intended for use from grade 4 to grade 9. There is no time limit, for it is the purpose of the test to examine the pupils' ability to reason in arithmetical situations.

Scoring the test.—The scores of the Stone Test are such as to serve three purposes: (1) survey, (2) supervision, and (3) teaching. For the first purpose, the papers are scored for answer only, inasmuch as this yields a single, comprehensive measure. Provision has also been made for the purpose of revealing the status of each pupil on each problem through the use of a diagnostic record sheet which is provided with the tests. The diagnostic scores will help the teacher to locate the source of individual difficulties, so that remedial work may be undertaken. For supervision, the diagnostic score is important, of course, in affording a basis for sound evaluation of the effectiveness of the instruction and of the program and methods employed. The test is provided with T-scores, age and grade norms, the latter of which are probably the most useful to the teacher. Statistical treatment of results has shown the test to have fairly high reliability, the highest reliability being in grades 5, 6, and 7, for which the test is chiefly intended and best suited.

⁸ This is an extension and improvement of the original Stone Reasoning Test in Arithmetic.

Function of the test.—The primary function of the test is to measure ability to solve arithmetical problems; that is, arithmetical reasoning. Speed of work is not a factor, inasmuch as no time limits are imposed. Although proficiency in the four fundamental operations is not directly tested, it is, nevertheless, a factor in the determination of group and individual status when the papers are scored for the correct answer only. But, of course, the test goes beyond the bare answers when used for analytic purposes, and proficiency in the fundamental operations is relegated to a position of unimportance. It is reasonable to say that the Stone Test succeeds rather well in measuring what it purports to measure.

It should be noted that a test of arithmetical reasoning is very frequently included in tests of "general intelligence."

FROM THE NEW STONE REASONING TEST IN ARITHMETIC:

1. Dorothy had 4 cents in her bank. Her uncle gave her 25 cents more. She bought a jumping rope for 20 cents. How much money did she have left? *Answer:.....*
13. A farmer pays \$4.20 to two boys for hoeing corn. Harry hoes 22 rows and George hoes 38 rows. How shall they divide the money? *Answer:.....*
21. One laborer can lay a pipe line in $5\frac{1}{5}$ days. Another laborer can do it in $7\frac{1}{3}$ days. How long will it take them to do the work together? *Answer:.....*

Comparison of tests.—The following table ⁹ includes important information regarding arithmetic tests. It will be noted, first, that in nearly every case the va-

⁹ We are indebted to Professor G. M. Ruch, who kindly supplied these data.

lidity is rather high; yet the coefficients are sufficiently far removed from unity or near-unity to warrant the interpretation that the tests do not all measure precisely the same functions.

VALIDITY, RELIABILITY AND OTHER DATA

(152 High Seventh and Eighth Grade Pupils, With a Mean IQ of 107)

								Norms	
	Valid- ity *	Relia- bility	<i>M</i> ₁	<i>M</i> ₂	<i>SD</i> _z	<i>SD</i> _y	Pa- pers scor- ed per hour	Grade Equiv- alent of Mean	Work- ing Time
Compass Survey	.78	.89	38.0	39.5	8.2	7.7	20	8 4	85'
Courtis "B"	.75	.87 **	26.8	26.8	11.1	11.4	20	H5	26'
Schorling-Clark-Potter	.79	.87	56.6	61.7	17.4	17.4	14	11	40'
Monroe	.76	.86	70.8	82.3	24.2	27.1	10	H6	26 1/4'
Stanford Reasoning	85								
Stanford Computation	70								
Total	78	.82	224.8	233.6	34.2	31.8	20	8 5	40'
Pittsburgh	.69	.65	32.0	32.4	3.7	3.2	41	H7	15'
Woody-McCall	.72	.50	30.1	30.0	2.8	2.9	36	Above H8 †	20'

* Validity defined as average intercorrelation with all other tests.

** Split-halves, stepped up by S-B (hence slightly too high).

† Norm for H8 is 28.5.

(Data by Eunice Adams and G. M. Ruch)

With the exception of the last two tests (Pittsburgh and Woody-McCall) the reliabilities are quite high and satisfactory. It is possible that one reason for the low reliability of the last two tests is their brevity, for they require only 15 and 20 minutes respectively.

The interesting thing about these data is the wide variation in the grade equivalents of the scores of the same group, the range being from the high fifth to the eleventh grade. This probably shows the untrustworthy character of some norms, for it is hardly likely that these discrepancies can be explained by the fact that they do not measure absolutely the same functions, or

by the fact that the training of the pupils has been such as to give them an advantage in some tests, while placing them at a disadvantage in others. These grade equivalents indicate the importance of studying the nature of the norms for any test which is to be used.

The logic of the situation suggests that a group of high seventh and eighth grade pupils toward the end of the year, and with a mean IQ of 107, should run close to 8.5 as a grade equivalent. Considering all the data of the table, it appears that the revised Compass and the Stanford norms are most trustworthy.

Conclusions.—From the brief survey of representative tests of arithmetic ability it should be clear that the technique of objective measurement has been applied to this subject with marked success. Several of the reasons for this are first, that the subject-matter lends itself to objective measurement, and second, that the importance of arithmetic as a school subject and as a necessary tool in daily life has been the incentive for considerable and varied study of the problem. As a consequence, the tests of arithmetic ability possess a very reasonable degree of reliability and validity; and they are so constructed as to afford the teacher an instrument which should increase the degree of her effectiveness in the teaching of arithmetic. This is so chiefly for two reasons: first, waste effort is eliminated, since the important elements of the subject and their relative difficulty have been indicated; and second, as with every other good test, weaknesses and strength of a pupil or class are revealed.

MATERIALS NEEDED

Compass Diagnostic Tests in Arithmetic—Grades 2 to 8. Tests 1, 2, 6, 7, 19 and 20, each 25 cents per 25; Tests 3, 4, 5, 8, 9, 10, 11, 12, and 16, each 50 cents per 25; Tests 13, 14, 15, each \$1.00 per 25; Tests 17 and 18, each \$1.25 per 25; teacher's manual, 20 cents. (Scott, Foresman & Company, Chicago, Ill.)

Courtis, S. A., Standard Research Tests in Arithmetic—Series B—for grades 3 to 8. Complete material for testing a class of forty pupils 75 cents; S. A. Courtis, 1807 East Grand Blvd., Detroit, Mich.

Courtis, S. A., Standard Practice Tests in Arithmetic, 48 graded lessons, two forms, A and B. For grades 4 to 8. Cabinet I, 576 lesson cards with guides for a class of 48, price \$8.50, Cabinet II, 288 lesson cards with guides for a class of 24, price \$6.50, Cabinet III, 144 lesson cards for a class of 12, price \$2.25. (World Book Company, Yonkers-on-Hudson, N. Y.)

Judd, C. H., The Cleveland Survey Arithmetic Tests for grades 3 to 8. (The Public School Publishing Company, Bloomington, Ill.) Sample set 10 cents, price \$1.90 per hundred.

Monroe, W. S., Diagnostic Tests in Arithmetic. For grades 4 to 8. (The Public School Publishing Company, Bloomington, Ill.) Sample set 15 cents, price 85 cents per hundred.

Monroe, W. S., Standardized Reasoning Tests in Arithmetic, forms 1 and 2. Test I is for grades 4 and 5, test II for grades 6 and 7 and test III for grade 8. (The Public School Publishing Company, Bloomington, Ill.) Sample set 8 cents, price 80 cents per hundred.

Stone, C. W., The New Stone Reasoning Tests in Arithmetic; Forms 1 and 2. Grades 4 to 9. (New York, Bureau of Publications, Teachers College, Columbia University.)

Woody, C. and McCall, Wm. A., Mixed Fundamentals, Forms I and II for grades 3 to 8. (New York Bureau of Publications, Teachers College, Columbia University. Sample set 10 cents, price 60 cents per hundred.)

SUPPLEMENTARY LIST OF TESTS

Buckingham Scale for Problems in Arithmetic.¹⁰

Grades 3 to 8.

Buswell-John Diagnostic Test for Fundamental Processes in Arithmetic (an individual test).¹⁰

All grades in which arithmetic is taught.

Clapp's Number Combination Tests.¹¹

Clapp-Young Arithmetic Test.¹¹

Grades 5 to 8.

Compass Survey Tests.¹²

Grades 2 to 8.

Denver Curriculum Tests in Arithmetic.¹³

Grades 2 to 8.

Foran Diagnostic Computation Scales.¹⁴

Grades 2 to 8.

Kinney Scales of Problems in Commercial Arithmetic.¹⁰

Lunceford Diagnostic Test in Addition.¹⁵

Primary Grades.

Monroe General Survey Scales in Arithmetic.¹⁰

Grades 3 to 8.

Otis Arithmetic Reasoning Test.¹⁶

Grades 4 and above.

Peet-Deerborn Progress Tests in Arithmetic.¹⁷

Grades 1 to 6.

Pittsburgh Arithmetic Scales.¹⁰

Grades 3 to 9.

Schorling-Clark-Potter Arithmetic Test.¹⁸

Grades 5 to 12.

¹⁰ The Public School Publishing Company, Bloomington, Ill.

¹¹ Houghton Mifflin Co., Boston, Mass.

¹² Scott, Foresman and Company, 625 S. Wabash Ave., Chicago, Ill.

¹³ Denver Public Schools, 414 Fourteenth St., Denver, Colo.

¹⁴ Catholic Education Press, 1326 Quincy St., N. E., Washington, D. C.

¹⁵ Bureau of Educational Measurements and Standards, Teachers College, Emporia, Kansas.

¹⁶ World Book Company, Yonkers-on-Hudson, N. Y.

¹⁷ Harvard Laboratory of Educational Psychology, Cambridge, Mass.

Stanford Achievement Test-Arithmetic.¹⁶

Grades 2 to 9.

Stevenson Problem Analysis (Arithmetic Reading Test).¹⁰

Grades 4 to 9.

Wildeman Standardized Test in the Fundamental Operations with Common Fractions.¹⁸

Wilson General Survey Test in Arithmetic.¹⁹

Grades 5 to 7.

Wisconsin Inventory Tests in Arithmetic.¹⁰

Grades 2 to 8.

SELECTED REFERENCES

Brown and Coffman. *The Teaching of Arithmetic* (Row Peterson, Chicago, 1925).

Buckingham, B. R., "Mathematical Ability as Related to General Intelligence" (*School Science and Mathematics*, Vol. 21, March, 1921, pp. 205-215).

Buswell, G. T. and John, L., "Diagnostic Studies in Arithmetic" (*Supplementary Educational Monographs*, No. 30; The University of Chicago Press, 1926).

Buswell, G. T. and Judd, C. H., "Summary of Investigations Relating to Arithmetic" (*Supplementary Educational Monographs*, No. 27; University of Chicago Press, 1925).

Douglass, H. R., "Development of Number Conception in Children of Pre-school and Kindergarten Ages" (*Journal of Experimental Psychology*, Vol. 8, 1925, pp. 443-470).

Freeman, F. N., *The Psychology of the Common Branches*. Chapter IX on Mathematics (Boston, Houghton Mifflin Company, 1916).

Howell, H. B., *A Fundamental Study in the Pedagogy of Arithmetic* (New York, The Macmillan Company, 1914).

Hunkins, R. V. and Breed, G. S., "The Validity of Arithmetical Reasoning Tests" (*The Elementary School Journal*, Vol. 23, p. 453).

Knight, Luse, and Ruch, "Problems in the Teaching of

¹⁸ Plymouth Press, 6749 Wentworth Ave., Chicago, Ill.

¹⁹ University Publishing Company, 2126 Prairie Ave., Chicago, Ill.

Arithmetic: A Syllabus for Discussion on Important Aspects of Elementary School Arithmetic" (Iowa Supply Company, Iowa City, Iowa, 1924).

Myers, G. C., *The Prevention and Correction of Errors in Arithmetic* (Chicago, The Plymouth Press, 1926).

Newcomb, R. S., *Modern Methods of Teaching Arithmetic* (Boston, Houghton Mifflin Company, 1926).

Stone, C. W., *Arithmetical Abilities and Some Factors Determining Them* (Contributions to Education. No. 19, Teachers College, Columbia University, 1908).

Thorndike, E. L., *The New Methods of Arithmetic* (Chicago, Rand, McNally & Company, 1921).

Thorndike, E. L., *The Psychology of Arithmetic* (New York, The Macmillan Company, 1922).

Wilson, G. M., "Preliminary Report on Arithmetic Reconstruction" (*National Education Association, Proceedings*, 1924, pp. 311-335).

Woody, C., *The Measurement of Some Achievements in Arithmetic* (Contributions to Education, No. 80. Teachers College, Columbia University, 1920).

CHAPTER X

GEOGRAPHY

The problem of measuring in geography.—The construction of a test in a content subject like geography is much more difficult than the construction of one in a formal subject like spelling, or even in arithmetic. The chief reason, perhaps, lies in the difficulty of determining what essentials shall be taught. Teachers and texts, for example, do not agree with respect to the relative amount of time and emphasis to be spent upon memory work and upon thought-material. Nor is there agreement regarding the relative importance of native geography, foreign geography, industries, form changes, knowledge and use of maps, ability to deduce and use geographical principles, and many others as well. Which of these and others should be emphasized, merely touched upon or ignored is a problem that cannot be answered in an arbitrary manner; nor can it be answered simply by means of a statistical analysis of the frequency with which topics occur in textbooks and in established curricula. To be sure, the statistical analysis aids in gaining an insight into prevailing practices; it does not, however, establish the wisdom of those practices. There enters here quite naturally, therefore, the question of values, even as in any and all school subjects and procedures.

Though we are not primarily concerned with prin-

ciples of curriculum building and selection, it is necessary to point out that the choice of a test in geography, or in a subject of the same type, must be in large part determined by the factors the teacher desires to measure and by the specified purpose of the test itself. One cannot simply select "a test of geography" for any given situation.

HAHN-LACKEY GEOGRAPHY SCALE

Description of the scale.—This is one of the oldest of the geography scales. It consists of a series of 217 questions arranged on the same general plan as the Ayres Spelling Scale; that is, the questions are arranged in twenty-three columns with from one to eighteen questions in each column. The questions in any column are of approximately equal difficulty, and the columns are arranged in order of difficulty. Norms for the different school grades are given at the top of each column. For example, a fifth grade class should make an average score of 88 in column "V" or a score of 92 in column "W" of the scale.

The questions in this scale are about equally divided between "memory ability" and "thinking ability." Those of the former type are printed in light face type while the latter appear in black face type. In order to make the scale diagnostic the author has classified the questions under seven headings. These are:

1. Knowledge of home geography.
2. Knowledge of the meaning of the technical terms or symbols.
3. Knowledge of the map as a geographical tool.

4. Ability to locate places in geography.
5. Ability to use constructive imagination to see geographical situations as they are.
6. Ability to think inductively or derive general principles.
7. Ability to think deductively or deduce geographical from general principles.

V	W	X	Y	
79	84	88	92	FOURTH GRADE
88	92	94	96	FIFTH GRADE
92	94	96	98	SIXTH GRADE
96	98	99	100	SEVENTH GRADE
96	98	99	100	EIGHTH GRADE
58. What direction do you live from the equator?	53. In what direction are you facing when your back is toward the north?	52. Name two animals used by the Eskimos.		
64. Name two important mountain ranges of the United States.	61. What is the capital of your state?			
11. Name the four seasons.	17. Name two things plants must have to live			
32. What is the direction half way between south and west?				
2. In what direction would you go to go to Canada?				
15. What is the capital of the United States?				
35. Name an animal useful to man in desert countries.				
55. Why is there not much farming done in Alaska?				
22. How do we know that there is air?				
4. Why is the Arctic Ocean not used much by sailors?				
48. How does the ocean help to furnish us food?				
56. Why are there more birds here in summer than in winter?				

Hahn-Lackey Geography Scale

DIRECTIONS FOR GIVING TESTS

1. Do not impose a time limit.
 2. Select from four to ten exercises from the list given in any one column. Write your selection of exercises for the test on the blackboard and proceed as in an ordinary school examination.

3. Give the following instructions to the pupils taking the test: "You will be given enough time to answer as many of these exercises as you can. Kindly read each exercise carefully to get its exact meaning, then write the best answer you can in the fewest words. Complete sentences are not necessary, words or phrases will do. You are not expected to be able to answer all the questions. Some of them will be very difficult or impossible. If you can answer the difficult ones, the credit due you will be that much greater. At any rate, please try hard to answer every exercise. Ask no questions about any of the exercises in the test. If your teachers should permit you to ask questions and then answer them for you, it would defeat the purpose of the test and your answers could not be used."

4. In scoring answers to exercises consisting of two or more parts, give credit for each part answered correctly. In general, give credit for an answer that clearly indicates a knowledge of the idea involved in the exercise. Copies of directions indicating specifically what to accept and what to reject in scoring answers may be obtained from the authors for Five Cents apiece.

5. The standards given are correct for 1922.

A SECTION OF THE EASIER END OF THE HAHN-LACKEY GEOGRAPHY SCALE

Method of using the scale.—The teacher may select a list of questions from one of the columns for use with

the class. The authors suggest that from four to ten questions should be selected. They may be memory or thought questions, as the teacher chooses. The pupils are allowed as much time as they need for answering. The answers, graded by means of a key, are marked on the basis of 100 as a perfect score. The grade for a pupil or the average grade for a class may be compared with what is to be expected by reference to the norms given at the top of the column from which the questions were selected.

Function of the scale.—A list of questions such as this scale offers provides the teacher with a profitable source of material for testing her pupils. But as a scale, it possesses certain features limiting its usefulness. For example, a wide selection of questions must be made from the scale, otherwise the likelihood is that those chosen will not measure what the teacher intends to measure. Questions almost entirely of the memory type may be chosen; and, if the teacher has been drilling her pupils in that type of material, the scores are likely to be spuriously high, and, therefore, misleading. On the other hand, the questions selected may be so chosen as to be unfair to the class. But if carefully employed and scored, this test is fairly satisfactory in testing knowledge and in helping to make an analysis of the difficulties or the qualities of geography achievement.

POSEY-VAN WAGENEN GEOGRAPHY SCALES

Description of the scales.—The Posey-Van Wagenen Scales are made up of six parts, four of which have two divisions, I and II, while the remaining two parts have

one division, II. In each scale, division I is for use in grades 5 and 6, while division II is intended for grades 7 and 8. Some of the scales deal specifically with the United States and North America; others deal with Europe, South America, Africa, and Asia. Other members of the series are general in nature. The questions, of which there are thirty of increasing difficulty in each scale, are of both types, thought and informational. "The material for these geography scales has been selected as representative of a wide range of interests. The selection of the material, the framing of the tasks, and their final arrangement into scales is the result of judgments based upon a wide range of geographical information and a wide experience in scale making and using." Four of the tests are for grades 5 and 6, and the others are for grades 7 and 8.

Method of using the scales.—The teacher selects the scale which is appropriate for the grade and subject-matter which she desires to test. The time limit of forty minutes, within which the pupils answer as many questions as possible, is sufficient to enable most pupils to answer as many items as they are capable of handling. Keys and methods of scoring are provided in the manual. The following norms, for the end of the school year, apply to any of the scales.

THE POSEY-VAN WAGENEN GEOGRAPHY SCALES

<i>School Grade</i>	5	6	7	8
<i>Norm</i>	68	74	80	86

Function of the scales.—No one of these scales covers a wide range of geographical knowledge, but each calls

for fundamental and representative material in geography. This is shown by the fact that the probable error is only 2.1 for a score on these scales. This means that in half the scores the pupil's actual ability in that phase of the subject tested will probably not vary from the scale scores by more than 2.1 points or units on the scale. This is approximately one-third of a grade difference. The scales are so constructed that a difference of 5 points between any two points on the scale is supposedly the same as five points difference in any other part of the scale. Also a score of 60, for example, represents the same value on any one of the scales for the fifth and sixth grades and may be compared with the score of a pupil on any of the seventh and eighth grade scales. In other words, the units are supposedly uniform throughout all the scales and directly transferable from one grade to another.

The scales are easy to give but the scoring is somewhat complicated. This feature of the scoring is necessary to make the comparisons between the different scales possible. While the separation of the scales into thought scales and information scales prolongs the testing program, it makes them of greater diagnostic value.

POSEY-VAN WAGENEN GEOGRAPHY SCALES

Information R (General), Division 1. Grades 5 and 6

GROUP II (Average value 69.5)

11.(65)

On what do each of these things grow: a vine, a bush, a plant, or a tree?

1. Olives
2. Dates
3. Figs
4. Rice

16.(70)

- (a) Is winter wheat sown in the spring, summer, fall or winter?
.....
- (b) Is it harvested in the spring, summer, fall, or winter?
.....

POSEY-VAN WAGENEN GEOGRAPHY SCALES

Thought S, Division 1. Grades 5 and 6

5.(59) Why does not the palm tree grow in Canada?
.....

6.(60) It costs very much less for the same weight of any material to be carried by boat than by train. At the same time the train takes less time to go between the same two places.
(a) If you had plenty of time and very little money, and had to go from New York to New Orleans, how would you go?
.....

9.(63) In which of these regions would you expect to find the largest number of people living: a fertile plain, a mountainous territory, an indented coast line open to the interior, a desert, an iron mining region?
.....

10.(64) It is much easier and cheaper to restore fertility to the soil by letting the land lie fallow or unused for a few years than to buy fertilizer or manure to put on it.
Is the land in sparsely settled regions more likely to be fertilized or allowed to lie unused?
.....

BUCKINGHAM-STEVENSON PLACE GEOGRAPHY TESTS

Description of the tests.—These tests consist of two parts, three forms of each being available. One part measures the pupils' knowledge of the location of cities, rivers, lakes, mountain ranges, etc., throughout the world; the other part is confined to the United States. The pupils are required, for example, to "name the continent on or nearest to which each of the following is located"; or, to "name the ocean into which each of the following rivers empty"; and other similar directions.

Method of using the tests.—The questions are to be read by the teacher and the answers written by the pupil. Sufficient time is allowed for all pupils to write the answers so that speed of writing is not a factor. Before the test is given, a preliminary trial exercise is studied to familiarize the pupil with the nature of the test. Standard scores are given for each form of the two tests as follows:

STANDARD SCORES (*United States*)

Grade	8th		7th		6th		5th		4th	
	High	Low								
Form 1, Score .	27	26	24	23	21	20	15	12	7	
Form 2, Score .	26	26	24	23	22	20	15	12	6	
Form 3, Score . .	29	28	25	24	23	21	15	12	7	

STANDARD SCORES (*World*)

Grade	8th		7th		6th		5th		4th	
	High	Low								
Form 1, Score . .	43	42	42	40	35	32	24	20	8	
Form 2, Score . .	43	43	43	38	36	33	26	20	11	
Form 3, Score . .	43	44	43	42	40	37	30	24	15	

Function of the tests.—These tests are rather satisfactory for the purpose for which they are intended. They measure the pupils' knowledge of geographical locations. They do no more than this; nor are they intended to measure other types of geographical information. The tests are simple and easy to give and score. Any teacher can give the tests without special preparation or material other than a single booklet containing the tests. They do not take up a large amount of time, and still they cover sufficient subject-matter to represent a fair sample of the pupil's ability in place geography.

COURTIS SUPERVISORY TESTS IN GEOGRAPHY

Description of the tests.—These measures are somewhat similar to the Buckingham-Stevenson Place Geography Tests, for they too test the pupils' knowledge of geographical locations, though in a different manner. There are two tests, with forms A and B of each; one is designed to measure knowledge of locations of oceans, continents, and countries of the world; the second is designed for locating states and cities of the United States. In each test a map is shown, the different parts being indicated by numbers. For example, in the map of the United States each state has a number. Test B consists of a list of 43 countries of the world to be located on the proper continents. After the name of each country, the pupil merely places the number of the continent upon which it is located. The same map is used for the identification of seven continents and five

oceans. The test for the United States is similar, except, of course, that states and important cities are to be located.

Method of using the tests.—The pupils are furnished copies of the test; and after a preliminary exercise, they are given one minute to answer as many as possible of the questions on continents and oceans, and then four minutes for the location of countries. For the United States Tests, four minutes are allowed for states and two minutes for cities. The score is a weighted per cent, with a maximum of 1,000 in each part of the test.

Function of the tests.—These tests are simple, brief, and easy to administer. Their use of maps probably adds to their value; but they are not so comprehensive as the Buckingham-Stevenson tests, and their usefulness, therefore, is restricted. However, statistical studies have indicated that the Courtis Tests do perform their limited function quite well. Their usefulness is further restricted, however, by the lack of norms or standards.

BRANOM'S DIAGNOSTIC TESTS IN GEOGRAPHY

Description of the tests.—This is the most complete set of tests in the subject, for they cover a very wide variety of topics and regions. In eight separate groups there are from two to four forms for testing each of the following: place, factual information, and problems. The eight groups are United States, North America, Europe, Asia, Australia, Africa, South Africa, and the World. The scope of the tests is evident when one

realizes the vast amount of information involved in questions of location and of fact, and in problems applying to nearly all parts of the known world.

Method of using the tests.—Of course, all divisions of the test cannot be utilized within one year's work. The proper tests are to be given on completion of the study of any region. The World test is designed for use when the course of study in geography for the grades has been completed. The tests are of the "true-false" and "multiple-answer" type. Each pupil receives a sheet and marks the correct answer or answers after each statement.

Function of the tests.—Obviously, these tests do not pretend to make merely a *limited* selection of the important topics for study in geography, for they are too nearly all-inclusive. It is their purpose to provide individual measures such as will furnish the teacher with a device to test the results of instruction in a given aspect of geography. They are further intended to serve the purpose of diagnosis, in order that instruction of individual or class may be better directed.

Conclusions.—From the brief discussion of the several tests included in this chapter, it seems apparent that their value depends in large part on how well they measure knowledge in those branches of geography which are regarded as essential. It may happen, of course, that some teachers and supervisors will regard the tests as limited in value because of the restricted nature of the materials included in most of them. In that view, it is not unlikely that they will be correct to a marked extent. Further, it may be felt that statistical studies have not shown the tests of geography to be as

highly reliable and valid as we desire.¹ From the statistical point of view, several of the tests (Buckingham and Courtis) bear up better than the others;¹ but they are measures which are quite factual and limited in scope. Yet, in spite of these limitations, wise use of these tests, and others, may serve the teacher well in effecting improved instruction and in giving her a better insight into the problems of the subject.

MATERIALS NEEDED

Bransom, M. E., Diagnostic Tests in Geography. For all grades where geography is taught, McKnight and McKnight, Normal, Illinois. Single copy, 1 cent; 80 cents per hundred; score sheet, 1 cent; key and norms, 20 cents.

Buckingham, B. R. and Stevenson, P. R., Place Geography tests. One tests "The World" and the other "United States." Three forms of each for grades 4 to 8. Only one copy of the test needed as the teacher dictates the test. The Public School Publishing Company, Bloomington, Ill. Booklet containing 3 forms each of both tests, 15 cents. Class record sheets, each 1 cent.

Courtis, S. A., Standard Supervisory tests—Geography—Location for grades 4 to 8. S. A. Courtis, 1807 East Grand Blvd., Detroit, Mich. Price \$1.50 for materials for a class of 40 children.

Hahn, H. H., Lackey, E. H., Geography Scale for grades 4 to 8. Only one copy needed. The Public School Publishing Company, Bloomington, Ill., price 20 cents; three or more, 15 cents each.

Posey, C. J., Van Wagenen, M. J., Geography Scales: Teachers Handbook 20 cents. Sample set, 30 cents, \$1.50 per 100. Scale Thought S Div. I (grades 5 and 6). Scale Thought R Div. II (grades 7 and 8). Scale Information R (gen.) Div.

¹ See Ruch and Stoddard, *Tests and Measurements in High School Instruction*, pp. 235 ff.

I (grades 5 and 6). Scale Information R (gen.) Div. II (grades 7 and 8). Scale Information S (gen.) Div. I. (grades 5 and 6). Scale Information S (gen.) Div. II (grades 7 and 8). Scale Information A (U. S. and North America) Div. I (grades 5 and 6). Scale Information A (U. S. and North America) Div. II (grades 7 and 8). Scale Information F (Europe) Div. II (grades 7 and 8). Scale Information K (S. A., Asia, Africa) Div. II (grades 7 and 8). The Public School Publishing Company, Bloomington, Ill.

SUPPLEMENTARY LIST OF TESTS

Forney Test in Map Reading Abilities.²

Gregory-Hagerty Geography Test.³

Grades 4, and 5 and 6.

Gregory-Spencer Geography Tests.³

Three upper grammar school grades.

Information-Problem Tests in Geography.⁴

Russell-Harr Geography Tests.⁵

Witham Standard Geography Tests.⁶

SELECTED REFERENCES

Courtis, S. A., "Measuring the Effects of Supervision in Geography" (*School and Society*, July 19, 1919).

Freeman, F. N., *The Psychology of the Common Branches. Geography*, Chapter VIII (Boston, Houghton Mifflin Company).

Lackey, E. E., "A Scale for Measuring Ability of Children in Geography" (*Journal of Educational Psychology*, October, 1918).

² Ginn and Company, 15 Ashburton Place, Boston, Mass.

³ Bureau of Administrative Research, College of Education, University of Cincinnati.

⁴ The Public School Publishing Company, Bloomington, Ill.

⁵ McKnight and McKnight, Normal, Illinois.

⁶ J. L. Hammett Co., Cambridge, Mass.

Posey, C. J. and Van Wagenen, M. J., "The Posey-Van Wagenen Geography Scales." *Teachers Handbook* (The Public School Publishing Company, Bloomington, Ill.)

Reed, H. B., *Psychology of Elementary School Subjects* (Ginn and Company, Boston, 1927).

Ruch, G. M. and Stoddard, G. D., *Tests and Measurements in High School Instruction*, Chapter 13 (World Book Company, Yonkers-on-Hudson, N. Y., 1927).

CHAPTER XI

HISTORY AND CIVICS

The problem of measurement in history.—In the construction of tests and scales, history presents the same type of difficulty as geography; for, as in geography, there are in history no universally recognized aims or objectives. The difficulty is enhanced by the fact that teachers of history regard as least significant those aims which could be most readily measured: namely, factual information and mastery of the text.¹ On the contrary, those aims ranked as most significant do not easily lend themselves to objective measurement, for they include such objectives as the "power of handling facts," "promotion of good citizenship," "development of discrimination." While it is likely that few, if any, hold the one extreme view that instruction in history should be devoted solely to the interpretation of human events, or the other extreme that only names, dates, and events should be acquired, it is true that measurement of the former is exceedingly difficult, whereas measurement of the latter alone would be regarded as inadequate.

It should be evident, however, that in history, as in any other subject, discrimination, reasoning, judgment,

¹ Koos, L. V., "The Administration of Secondary School Units." Supplementary Educational Monographs, Vol. I, No. 3, University of Chicago, 1917.

constructive imagination, citizenship, etc., can not be promoted in a vacuum; there must be materials—that is, facts—with which to reason and form judgments. Thus, it may well be that the author of a test in history or civics is justified in starting first with the essential information. Here again, another question arises. What is the essential information? But this is a question for the teacher and historian to deal with. It will be the task of the teacher to select the tests which in her judgment make the closest approach to her aims and objectives.

History

HAHN HISTORY SCALE

Description of the scale.—This scale is constructed on the same general principles as the Hahn-Lackey Geography Scale. It consists of about 275 questions arranged in columns. The questions in any column are of approximately equal difficulty. Standard scores for the eighth grade are given at the top of each column. The scale is also for use with the seventh grade, and following each question is a number giving its value in a seventh grade test.

The scale is supposedly diagnostic in that the questions have been classified under nine different abilities. These range from memory ability for historical facts to ability to see connections and make historical comparisons and judgments.

Method of using the scale.—The teacher in the eighth grade should “select from four to ten exercises from any one step” on the scale. These exercises may be writ-

STEPS	A	B	C	D
MAY STANDARDS EIGHTH GRADE	0	1	2	4
	137 What was President Wilson's plan of solving the tariff question? (0)		9 9 7 4 1 0 1 2 4 6 8	17 1 2 4 6 12

The Hahn History Scale

DIRECTIONS FOR GIVING TESTS

1. Do not impress a time limit.

2. Select from the 100 exercises from the list given in any one step. Write your selection of exercises on the blackboard and proceed as in an ordinary school examination.

3. Give the following instructions given to pupils in the preliminary test: "You will be given enough time to answer as many of these exercises as you can. Kindly read each exercise carefully and get the meaning." Then say: "THE BEST ANSWERS IN THE TEST" You know. Complete sentences or statements are not necessary words or phrases will do. We do not expect you to be able to answer all the exercises. Some of them were made difficult on purpose. If you can answer one or two of the exercises in the test, you have done well. At any rate please try hard to answer every exercise. Kindly ask no questions about any of the exercises in the test. If your teacher should permit you to ask questions and then answer them for you, it would defeat the purpose of the whole test and your answers could not be used."

DIRECTIONS FOR SCORING ANSWERS

In scoring answers to exercises consisting of two or more parts allow credit for each part answered correctly. In general, allow credit for any answer which is correct in meaning, regardless of the exact words used. Copies of directions indicating specifically what to accept and what to reject in scoring answers may be obtained for Five Cents apiece.

26 Give one cause why England took new interest in America in English Civil War. (1)

27 Why did the British hold the northern and western posts after the Revolutionary War? (2)

28 Why did states pass after Missouri was admitted? (0)

29 Give two causes of the terms of President Jackson (1828-1836) long steps were taken toward democratic government in the United States but also in England and France. (0)

30 What evidence can you give to show that our institutions are becoming more democratic? (1)

31 What was the Portsmouth Peace conference? (0)

32 Explain the serious trouble which arose over the count of electoral votes in 1876 not possible today? (0)

33 On what point did Virginia and England agree during the days of James I and Charles II? (4)

34 How did the colonial legislatures manage to hold the colonial governors, with whom they had many disputes, in check? (0)

35 Explain the reasons why Jackson opposed the United States Bank. (1)

36 What event taught Madison the danger of going to war unprepared? (0)

37 What doctrine did Keyes set forth in the Webster-Hayne debate? (0)

38 Name three acts of Congress that increased the trouble between the North and the South with reference to slavery. (0)

39 How did Madison try to bring England and France to terms? (1)

40 Name two of the most important problems which the Jacksonian democracy had to face. (0)

41 Name the first state to come to the United States in increasing numbers between 1840 and 1860? (1)

42 Why was not the Compromise of 1850 able to stop the division between the North and the South? (0)

43 Name two occasions prior to 1860 when state sovereignty was ad vocated. (1)

44 Tell in the blanks. Our greatest grievance as a result of the war between the United States and France (1812-1815) was the _____ by the _____.

45 Under the final plan of Congress what did the people of a seceded state have to do to get back into the Union? (1)

46 What act of Congress brought about a better feeling between the North and the South in 1877? (1)

47 What was the main point made upon the United States with reference to Japanese school children in California? How was this trouble settled? (1)

48 Name two important events in our financial history since 1877. (1)

49 What are the two main propositions taken today in regard to the tariff? (0)

50 What was the purpose of the Pan American Congress? (0)

A SECTION OF THE HAHN HISTORY SCALE

ten on the blackboard and the pupils given all the time they need in answering the questions. The papers are to be graded by the usual method of scoring on the basis of 100 points for a perfect score. Correct answers

to each question are given in a key which is furnished with the scale.

The seventh grade teacher may select either a list of questions with the same seventh grade values or a list with different values and average these values for the seventh grade standard. For example, if four questions have seventh grade values—as given after each question—of 38, 40, 40 and 46, a seventh grade class should make an average score of 41 on this test.

Function of the scale.—In general the same criticism may be made of this scale as of the Hahn-Lackey Geography Scale. It provides the history teacher with a large list—probably not a “complete” list as the author suggests—of fairly representative questions for use in seventh and eighth grades. The material from which the questions were taken is common to six modern texts in history. The diagnostic feature of the test gives the teacher an opportunity to determine in which types of history the pupils are well prepared and in which types they are deficient. No attempt has been made to indicate the relative importance of the several kinds of learning in history. While the scale is not too difficult for use by the class-room teacher if she will follow the directions carefully, it seems likely that the same material would be more usable if made up into a series of alternative tests with definite norms.

BARR DIAGNOSTIC TESTS IN AMERICAN HISTORY

Description of the tests.—Recognizing the fact that the content of history is not standardized, Barr has constructed a test for each of five categories in history, which are:

1. Comprehension of historical facts.
2. Chronological judgment.
3. Historical evidence.
4. Time relation of events.
5. Causal relationship in history.

Following a practice test are the five series of tests. The first consists of paragraphs chosen from various sources of American history, followed by sets of questions. In the second, lists of persons or events are to be arranged in chronological order. In the third, various sources of history are to be weighed as to their importance or reliability. In the fourth, series of historical events are to be arranged in order of importance. In the fifth, historical events are to be connected with their causes.

Method of using the tests.—Six minutes are allowed for each test. There are two forms of the tests, 2A and 2B, of which 2A is the more difficult.

Each question answered has a certain weight or score value. This value is given after the answer to the question on the scoring sheet. A pupil's score is the sum of the weights of all the questions answered correctly. These scores are kept separately for each of the five categories.

Function of the tests.—These tests present an interesting and suggestive attempt to solve the problems of tests in history. The separation of the material into five groups should give them diagnostic value and thereby make possible a differentiation of teaching methods to meet individual or class differences. The materials of the tests do not in any sense cover the whole field of American history; but that which has

been chosen is significant and representative. Again, some historians might differ from the author with respect to the answers to certain questions, but in general the answers are unequivocal. The several parts of the test call for varied and complex abilities, ranging from the almost purely informational to tests of judgment depending more upon "general intelligence" than upon historical information.

As a standardized measure, however, the Barr Tests are not finished; in fact they are not "standardized." The tests are to be commended for their suggestive value and for indicating certain objectives in the teaching of history.

SAMPLE PAGE FROM THE BARR DIAGNOSTIC HISTORY TESTS

TEST III

7. *Following is a letter from Governor Gibson:*

*Vincennes, Indiana Territory,
August 10, 1812.*

"Col. Wm. Hargrave,

Commanding Mounted Rangers:

"Two scouts from this post were at a point on the west White river thirty miles east of the forks and saw two old Delaware Indian men who have a lone wigwam at that place. These Indians were friendly and have been for a long time. They said that several Pottawattamies had recently been at that point and told them—'soon we will go to the Ohio river —get heap horses—maybe get scalps—the British drive Americans away soon.'

John Gibson,
Acting Governor."

Put a cross (X) before each of the following questions that you would like to have answered if you were writing

a history of the War of 1812 and came upon the above account.

- (a) What was the training and profession of the writer?
- (b) Was the writer prejudiced?
- (c) Was the author's literary style good?
- (d) Was the writer in a position to know the facts?
- (e) Did the writer take the trouble to get the facts?

End of Test III. Check your answers and then wait quietly until all finish.

TEST IV

1. Put a cross (X) before the event in the following list which has been of the greatest importance in American History.
 - (a) Braddock's defeat
 - (b) Burr's conspiracy
 - (c) The Hayes-Tilden contest
 - (d) The discovery of America
 - (e) The Webster-Hayne debate
2. Put a cross (X) before the event in the following list which has been of the greatest importance in the economic development of the United States.
 - (a) The Tariff Act of 1832
 - (b) The invention of the telephone
 - (c) The panic of 1873
 - (d) The laying of the Atlantic cable
 - (e) The introduction of railway transportation

HARLAN TEST OF INFORMATION IN AMERICAN HISTORY

Description of the test.—This test consists of ten exercises for measuring different types of historical

knowledge. Each exercise, except No. IV, contains from four to six questions dealing with men, events, and dates in American history. Exercise IV contains two questions in civics.

Method of using the test.—The pupil is allowed as much time as necessary to finish the test. Most pupils finish in twenty-five minutes.

The scoring is done with the aid of a scoring key. Each element of each exercise receives a score of 2, 1, or 0, according to whether it is correct, half correct, or entirely wrong. There are fifty elements in the test, and 100 constitutes a perfect score. Median scores for the end of the school year are given as follows:

HARLAN AMERICAN HISTORY TEST

<i>School Grade</i>	7	8
Norm	56	86

Function of the test.—This test is simpler than either of the two previously described, and naturally so, inasmuch as its scope is limited to information only. The subject-matter is apparently well selected as indicated by the fact that it was based upon the Bagley and Rugg² study of the content of twenty-three standard text books in American history. The test would, however, be more valuable if there were several alternative forms.

² Bagley, W. C., and Rugg, H. O., "The Content of American History as Taught in the Seventh and Eighth Grades," (*University of Illinois, School of Education Bulletin*, No. 16, 1916).

FROM THE LAST PAGE OF THE HARLAN TEST OF
INFORMATION IN AMERICAN HISTORY*EXERCISE VIII. Score*

Below are some general statements concerning the history of our country. Prove that they are true by stating a typical example or instance in American History which has shown them to be true.

1. *One method employed by a nation in acquiring territory is by conquest.*
.....
2. *The final decision of civilized people is that the enslavement of one people by another is wrong.*
.....
3. *The national congress has regarded unrestricted immigration as dangerous to the welfare of the nation.*
.....
4. *An exaggerated idea of the power of the president has, at times, endangered the life of the president.*
.....

EXERCISE IX. Score

The following topics represent matters of importance in the history of the United States. State definitely of what significance each has been.

1. *Articles of Confederation.*.....
2. *Mason and Dixon's line.*.....
3. *Monroe Doctrine.*.....
4. *The Tariff.*.....

VAN WAGENEN AMERICAN HISTORY SCALES

(Revised Edition)

Description of the scales.—These scales are constructed on the same general principle as the Posey-Van Wagenen Geography Scales. There are information scales for grades 5 to 12, and one thought scale for grades 7 and 8. The information scales consist of series of questions in American history, some of which are to be answered by checking the right answer from a list of possible responses, others by writing the correct answer. There are thirty questions, and they cover a wide range of historical information, while the several scales combined cover the whole period of American history. The thought scale consists of lists of historical facts or events followed by a question regarding the cause or reasons for the fact or event. The answer to the question is not given in the material read but is to be deduced from the material given plus the pupil's wider experience with historical facts and life. The answers are a matter of judgment on the part of the pupil. There are also thirty questions in this scale.

QUESTIONS SELECTED FROM THE VAN WAGENEN AMERICAN HISTORY SCALES—REVISED EDITION

*From INFORMATION. R. General Division 1. Grades 5 and 6.
Group I*

3. (56) Put a check mark in front of each of these things which were in use during the Civil War.

- Submarine
- Poison Gas

- Cavalry
- Ironclad war vessels
- Aëroplanes

4. (57) What were the first four European countries to make settlements in America?

- 1
- 2
- 3
- 4

From THOUGHT. R. For Grades 7 and 8. Group II

12. (77) Previous to the Civil War a large part of the Southern cotton crop was exported to England.

What was evidently one of the chief occupations of England?

13. (78) In 1800, Spain gave Louisiana up to France. The United States, fearing that France might set up a colony and control the Mississippi River, was anxious to get Louisiana. In 1803, Napoleon of France feared that Great Britain was about to seize his American territory.

What would you expect Napoleon to do?

14. (79) In 1810, nine tenths of our foreign trade (980,000 tons) was carried in American vessels. The War of 1812-14 stopped the importation of foreign-made goods.

In what industry would you expect American capital soon to have become invested?

Method of using the scales.—The pupils are furnished with copies of either the thought scale or one of the information scales. Both, of course, may be used. Forty minutes are allowed for the test.

The method of scoring the test, though clearly illustrated in the manual, is rather complex. The author of

the test claims that his method yields comparable units, so that a difference in one part of the scale is directly comparable with differences in any other part. This factor is interesting but of doubtful importance in a test of this sort where the materials are not of proved validity.

Function of the scales.—The Van Wagenen scales represent historical items which are regarded as acceptable and significant, for they were checked and criticized by historians. But the scales place by far the major emphasis upon information, for of the twenty forms already issued or about to appear, nineteen test information. Although the factor of comprehension is not ignored, the answers depend to a very noticeable extent on memorized material.

PRESSEY-RICHARDS TESTS IN THE UNDERSTANDING OF AMERICAN HISTORY

Description of the tests.—There are four parts to this test: (1) character judgment; (2) historical vocabulary; (3) sequence of events; and (4) cause and effect relationships. Each part consists of twenty-six items, one being a practice exercise.

SAMPLE EXERCISES FROM THE PRESSEY-RICHARDS TESTS FOR THE UNDERSTANDING OF AMERICAN HISTORY

Test I—Character Judgment

Complete directions instruct the pupils to underline the one word after each man's name which he thinks best describes him.

14. Aaron Burr: conscientious, honored, shy, disloyal.
15. Daniel Webster: Eloquent, quarrelsome, clever, dominating.
26. Herbert Hoover: Efficient, assertive, nervous, talkative.

Test II—Historical Vocabulary

The pupils are directed to underline the one of the four statements after each question which is the correct answer.

4. What is a confederacy? A disunion, A colony, A commonwealth, A league of states.
5. What is an autocracy? Representative government, Mob law, Self-government, An absolute form of government.
22. What is a panic? A mass of people, A political disturbance, A financial crisis, A gold rush.

Tests III—Sequence of Events

Of the four events in each question the pupils are directed to underline the event that happened the longest time ago.

5. First Continental Congress, Hartford Convention, Constitutional Convention, Declaration of Independence.
6. Battle of Yorktown, of Saratoga, of Bunker Hill, of Lexington.

Test IV—Cause and Effect of Relationship

Of the four events given in each question three are causes and one the effect. The pupils are directed to underline the effect.

10. Ratification of the Treaty of 1819, Holy Alliance, Need for independence in South America, Monroe Doctrine.
11. Issuing of paper money by state banks, Closing of U. S. Bank, Specie Circular, Panic of 1837.

The Character Judgment test contains the names of prominent persons or groups in American history. Following each name there are four adjectives. The pupil is told to underline the adjective after each name which best describes the person or event. The Historical Vocabulary test contains twenty-six historical terms. Each is followed by four possible answers. The pupil is directed to underline the correct answer to each question. The Sequence of Events test consists of twenty-six sets of four historical events. The pupil is directed to underline the event in each set that happened the longest time ago. The Cause and Effect Relationship test consists of twenty-six lists of events, with four in each list, three of which acted as "causes" and one of which was the "effect." The pupil is asked to underline the "effect" in each list.

Method of using the tests.—Five minutes are allowed for the first test, six minutes for the second, six for the third and eight for the last. One point credit is allowed for each of the twenty-five exercises in each test, exclusive of the practice exercise. The total possible score is, therefore, 100 points. Norms for the test are given as follows:

NORMS FOR THE PRESSY-RICHARDS TESTS FOR THE UNDERSTANDING OF AMERICAN HISTORY

<i>Grades</i>	6	7	8	12
<i>Total Score</i>	21	29	41	63
Test I	6	7	10	15
Test II	5	7	11	17
Test III	5	8	11	16
Test IV	5	7	9	15

Function of the tests.—This test, simple and easy to administer, is designed to measure the abilities listed by it. The authors do not, however, explain the selection of their material, so that the validity of the test will depend, in the eyes of the teacher, upon how significant are the included historical facts. This is true even if it be granted that the four listed aspects of the test are foremost in importance. Reference to the Twenty-First Yearbook of the National Society for the Study of Education will show that there are items in the Pressey-Richards test which are regarded by some as unimportant. Statistical treatment³ of the test, however, has indicated it to be among the most reliable of the history scales.

COLUMBIA RESEARCH BUREAU AMERICAN HISTORY TEST

Description of the test.—Tests for secondary schools have been increasing in number within recent years. This test, one of the most recent, is intended for use with high school and college students. It consists of four parts: (1) eighty true-false statements covering the period from colonial times to the present; (2) matching fifty historical items of various sorts; (3) fifty multiple-choice questions; and (4) twenty completion sentences. Throughout, "the test is designed to be representative of the various aspects of American history which are approved by the best available authorities."

³ Ruch, G. M., et. al., *Objective Examination Methods in the Social Studies*, Chapter VI. (Chicago, Scott, Foresman and Company, 1926).

Method of using the test.—The test may be given in two sittings, but it is desirable to give it in only one, if possible. The test is long, having a time limit of an hour and a half. In part one, the score is the number right minus the number wrong;⁴ in the remaining three parts, the score is the number of right responses.

Function of the test.—The test demands a rather thorough grounding in American history, chiefly of the factual type. The authors maintain, however, that the test is not merely a measure of memorized facts, but “of ability to make sound judgments and penetrating inferences from concrete facts.” The test very likely does this, but to a limited extent; for it may well be doubted whether a standardized test of the true-false, multiple-choice, or completion type is able to measure these functions to more than a limited degree.

The materials of the Columbia Research Test seem to be carefully selected for their purpose and rather well validated. They show a high index of reliability. Undoubtedly they are suggestive for the teacher of secondary school history, just as those previously described are suggestive chiefly for the elementary school teacher.

FROM THE COLUMBIA RESEARCH BUREAU AMERICAN
HISTORY TEST

Part II

DIRECTIONS. Below are eight groups of items, each of which is divided into two columns. Each item in the left-hand

⁴ The “right-minus-wrong” method of scoring is intended theoretically to compensate for mere guessing of answers. There are, however, certain psychological objections to the method, which make its use unjustifiable at times.

column is numbered. Each item in the right-hand column is followed by parentheses. Place in the parentheses the number of that item in the left-hand column that is associated with the item in the right-hand column. Each group is a separate problem; do not match items in different groups. *Twenty minutes.*

SAMPLES.

a. 1. 1492 Declaration of Independence ..(3)
2. 1620 Discovery of America(1)
3. 1776

I. 1. Pennsylvania	First permanent settlement in America	()
2. Massachusetts	Tobacco	()
3. New York	Largest number of German settlers	()
4. Virginia	Samuel Adams	()
5. Georgia	Robert Morris	()
	Last colony to be established	()
	Rum manufacture	()
	Poor Richard's Almanac	()
	Zenger Trial	()
	Dutch West India Company ...	()

Part III

DIRECTIONS. Below are several statements and questions, each of which is followed by five phrases. Mark in the parentheses the number of that phrase that correctly completes the statement or answers the question. (One, and only one, phrase is correct in each case.) *Thirty-five minutes.*

SAMPLE.

a. One of the principal products of colonial New York was—
1 rice 2 indigo 3 flour 4 gold 5 aluminum ... (3)

1. The five Intolerable Acts were authorized by—
1 the Colonial Assembly of Massachusetts 2 the
First Continental Congress 3 the royal Governor
of Massachusetts 4 the Second Continental Con-
gress 5 the British Parliament()
2. The bulk of intercolonial commerce was carried by
means of—
1 canals 2 inclined railways 3 pack horses
4 stagecoaches 5 river and coastwise boats()
3. The Molasses Act of 1733 was designed to aid—
1 English West Indian planters 2 colonial mer-
chant shippers 3 French sugar growers 4 Eng-
lish merchants 5 Dutch carriers.....()

Civics

BROWN-WOODY CIVICS TEST

Description of the test.—This test, intended to measure objectively the pupils' achievement in civics, is based principally upon subject-matter of the high school, though it may also be used in grades 7, 8, and 9. Each exercise of the test is based upon materials "common to at least five of nine of the most widely used textbooks in Civics, which were carefully and minutely analyzed preliminary to the construction of the test." For the authors of the test, therefore, the aims and objectives of the study of civics are those which have the sanction of wide current practice.

There are three parts: (1) civic vocabulary, of the multiple-choice type; (2) civic information, in the form of the "yes-no" type; and (3) civic thinking, in multiple-choice form, in which a problem is stated, and the pupil must select the answer or appropriate reason.

Method of using the test.—Each pupil receives a booklet containing all three parts. The time limit is thirty-five minutes. The scores for parts I and II are the number correct in each; in part III the number correct is multiplied by 3; that is, it is a weighted score to give it greater importance in the total score.

FROM THE BROWN-WOODY CIVICS TEST

Part I. Civic Vocabulary

DIRECTIONS. Draw a line under the word, or group of words, in parentheses, which, in a civic sense, most nearly means the same as the first word in each line.

Begin here.

1. statue — (law, constitution, tradition, custom) 1
2. thrift — (stinginess, riches, greed, economy) 2
3. coöperation — (unselfishness, working-together, harmony, without friction) 3
4. urban — (pertaining to travel, pertaining to country, pertaining to city, pertaining to street cars) 4
5. treason — (punishment, defeat, betrayal, defiance) 5

Part II. Civic Information

DIRECTIONS. Draw a line under the right answer to each question.

Begin here.

1. Is the United States a democracy? Yes No 1
2. Is the Constitution of the United States the highest law of the land? Yes No 2
3. May any adult become a candidate for office, local or national? Yes No 3
4. As a general rule does ignorance of the law excuse its violation? Yes No 4
5. Is the President of the United States the executive officer of the nation? Yes No 5

Part III. Civic Thinking

2. Some of the members of your community are interested in securing a park for the use of the children in the west end of the city. You live on the east side of the city and would profit little, if at all, if the proposition should carry at the election. You are fully aware that a large sum of money is needed to purchase the property and that you must assume your share of the financial burden by paying a higher rate of tax on the property which you own. Why should you vote for the proposition?

1. Your property will increase in value.
2. Friends will benefit if the issue carries.
3. You may want to use the park.
4. It will entail but slightly higher taxes.
5. The park will benefit the community.

Function of the test.—This test measures specific information—including a specialized vocabulary—and the application of certain civic principles to specific situations. Here, as in the case of history tests, one's estimate of the material will depend upon how representative and essential he believes it to be. There can be little question, however, that the items are based upon what appears to be the most widely accepted practices in the subject. Still it is doubtful whether we are able at present to measure the broader and more fundamental purposes of teaching in civics: namely, interest in and awareness of social needs and responsibilities.

As a matter of technique, this test is to be criticized because of the demands it makes on reading ability.

Conclusions.—We may summarize the discussion of tests in history and civics by indicating their principal

difficulties and contributions. The tests measure information for the most part, although it is possible to go beyond mere factual data by including materials in the form of problems requiring judgment and discrimination. The importance of the materials included may be questioned; yet it is reasonably safe to say that the standard test will present a greater percentage of significant items than the tests of individual teachers. This is no reflection upon teachers; it is merely a fact inherent in the different methods of examination construction employed by teachers and makers of tests. Though tests of history and civics have not yet reached the same level of development as those of arithmetical operations, spelling, and reading; they do, nevertheless, make possible the better measurement of the effectiveness of teaching in history and civics; and, in addition, they contribute to the formulation of the aims and objectives to be achieved in these subjects.

MATERIALS NEEDED

Barr, A. S., Diagnostic tests in American History, Series B for 8th grade and senior high school. (The Public School Publishing Company, Bloomington, Ill.) Sample set 15 cents. Price \$4.00 per hundred.

Brown-Woody Civics Test. Forms A and B; for grades 7 to 12. (World Book Company, Yonkers-on-Hudson, N. Y.) \$1.30 per 25.

Columbia Research Bureau American History Test, Forms A and B; for high school and college. (The World Book Company, Yonkers-on-Hudson, N. Y.) Specimen set, 30 cents; 25 booklets, manual of directions, class record and scoring key, \$1.50.

Hahn, H. H., The Hahn History Scales for grades 7 and 8. (The Public School Publishing Company, Bloomington, Ill.)

One copy sufficient. Price 20 cents; three or more copies 16 cents each.

Harlan, C. L., *Test for Information in American History* for grades 7 and 8. (The Public School Publishing Company, Bloomington, Ill.) Sample set 6 cents, price 80 cents per hundred.

Pressey, L. W., and Richards, R. C., *A Test for the Understanding of American History*, for grades 6, 7 and 8, and senior high school. (The Public School Publishing Company, Bloomington, Ill.) Sample set 10 cents, \$2.00 per hundred.

Van Wagenen, M. J., *Reading Scales, History*, for grades 5 to 12. (The Public School Publishing Company, Bloomington, Ill.) Scale R Information (General) Division 1 for grades 5 and 6. Scale R Division 2 for grades 7 and 8. Scale Thought R Division 1 for grades 5 and 6, Division 2 for grades 7 and 8. Bureau of Publications, Teachers College, Columbia. \$2.00 per hundred; thought scale, \$2.25 per hundred.

SUPPLEMENTARY LIST OF TESTS

Almack Tests in American Civics and Government.⁵

American Council Civics and Government Test.⁶

American Council European History Test.⁶

Burton Civics Test.⁵

Denny-Nelson American History Test.⁶

Denver Curriculum Semester Tests in American History and Government.⁷

Denver Curriculum Tests in World History.⁷

Gregory Tests in American History.⁵

Grades 7 to 12.

Gregory-Owens Test in Mediaeval and Modern History.⁵

High school and normal school.

⁵ Bureau of Administrative Research, College of Education, University of Cincinnati.

⁶ World Book Company, Yonkers-on-Hudson, N. Y.

⁷ Denver Public Schools, 414 Fourteenth St., Denver, Colo.

Hill Tests in Civic Information and Attitudes.⁸
Hill-Wilson Civic Action Test.⁸
Kepner Background Test in Social Sciences.⁹
Public School Achievement Test in History (Orleans).⁸
Sloyer Test in World History.¹⁰
Tyrell American History Exercises.¹⁰
Vannest Diagnostic Test in Modern European History.¹¹
Van Wagenen Reading Scales in History.⁸
Witham Comprehensive History Tests.¹²

SELECTED REFERENCES

Bagley, W. C., and Rugg, H. O., "The Content of American History as taught in the Seventh and Eighth Grades" (*University of Illinois, School of Education, Bulletin No. 16*, 1916).

Brinckley, S. G., *Values of New-Type Examinations in the High School, with Special Reference to History* (Contributions to Education, No. 161, Teachers College, Columbia University, 1924).

Committee on Social Studies of the Commission of the Re-organization of Secondary Education of the National Education Association, "Social Studies in Secondary Education" (*U. S. Bureau of Education Bulletin*, No. 28, 1916).

Hardy, R. E., "New Types of Tests in Social Science" (*Historical Outlook*, Vol. 14, Nov., 1923, pp. 326-328).

Kepner, P. T., "A Survey of the Test Movement in History" (*Journal of Educational Research*, Vol. 7, April, 1923, pp. 309-325).

Koos, L. V., "The Administration of Secondary School Units" (*Supplementary Educational Monographs*, Vol. 1, No. 3, University of Chicago, 1917).

Latshaw, S., "History Tests and Scales" (*High School Journal*, Vol. 6, pp. 13-15, 1923).

⁸ The Public School Publishing Company, Bloomington, Ill.

⁹ Harvard University Press, Cambridge, Mass.

¹⁰ The Palmer Company, 120 Boylston St., Boston, Mass.

¹¹ Indiana University Bookstore, Bloomington, Indiana.

¹² J. L. Hammett Company, Cambridge, Mass.

Ruch, G. M., et al., *Objective Examination Methods in the Social Studies* (Scott, Foresman and Company, Chicago, 1926).

Starch, D., *Educational Psychology* (Revised), Chapter XXII (New York, The Macmillan Company, 1927).

Tryon, R. M., *Teaching History in the Junior and Senior High School* (Scott, Foresman and Company, Chicago.)

Twenty-second Yearbook of the National Society for the Study of Education, Part II (The Public School Publishing Company, Bloomington, Ill., 1923).

Van Wagenen, M. J., "Revised Van Wagenen History Scales" (*Teachers College Record*, Vol. 27, No. 2, pp. 142 ff., 1925).

CHAPTER XII

MUSIC AND DRAWING

Music as a special talent.—Unlike most other school subjects, musical ability appears to be rather specialized. Among most subjects, there are many elements in common, and it has been demonstrated that the correlation between ability in the usual subjects is rather high.¹ That is, if a pupil achieves high marks in one subject, it is more likely than not that he will do as well, or nearly as well, in others. Likewise, those pupils doing mediocre or poor work in one subject will most likely be at approximately the same level in others. On the contrary, musical capacity *may* have little relation to capacity for other materials. This does not imply that success in most subjects carries with it failure in music, or *vice versa*, for in that event we should have an inverse relationship, or negative correlation.² The facts are simply that there is no *necessary* relationship between “general” ability and musical ability. It is possible for otherwise poor pupils to perform creditably in music, and for otherwise superior individuals to be deficient in musical ability and appreciation. This fact is demonstrated by the very creditable orchestras and

¹ Hollingworth, L. S., *Special Talents and Defects* (New York, The Macmillan Company, 1923). Also Starch, D., “Correlation Among Abilities in School Studies” (*Journal of Educational Psychology*, Vol. 3, pp. 415-418, 1913).

² See Chapter XVIII for an explanation of correlation.

bands of inmates of institutions for the feeble-minded. It should be borne in mind, however, that in the field of music, as in many others, the individuals of outstanding excellence are persons of superior quality, such as Bach, Schubert, Wagner, Paderewski, Kreisler, and others, including outstanding conductors of orchestras; for genuine creative effort and valid interpretation in all fields of knowledge require a high order of intelligence. There are exceptions, of course, but it is the nature of the vast majority of mankind which determines our "generalizations," though it is necessary to be wary of generalizations in dealing with any given individual. But by and large, it has been demonstrated that musical capacity shows little or no functional relationship with "general" capacity. This fact is in part due, no doubt, to the need in music for such specific abilities as pitch discrimination, perception of intensity, sense of time, etc., all of which may be poor in a person without affecting his "general intelligence."

Musical capacity exists to some degree in most persons. Knowledge of music depends upon the amount and quality of training received, conditioned by the capacity with which the individual starts. Some musical instruction is given in nearly all schools; in the better systems this instruction is frequently of a high order. It is desirable to know, therefore, what results are being achieved; to discover, if possible, which individuals should receive special encouragement and training in music; and, in general, to determine the nature of musical instruction to be given the individual or the group.

The foregoing discussion is for the most part true of drawing.

The tests about to be described, while by no means altogether satisfactory or complete, are very useful instruments and represent the nature and extent of measurement in music and drawing at the present time.

Music

SEASHORE MEASURE OF MUSICAL TALENT

Description of the tests.—These tests of "native" musical ability consist of a series of six phonograph records especially constructed for this purpose. The first record is a test in pitch discrimination. One hundred pairs of tones are sounded by playing this record on an ordinary phonograph. The pupil or class is furnished with score sheets on which they are to make a judgment as to whether the second tone in each pair is higher or lower in pitch than the first of the pair. The first series of ten pairs of tones varies by 30 vibrations from the standard which is about A \sharp (435 vs.). This difference is apparent to almost everybody of school age or above. The sixth series of ten tones differ by only $\frac{1}{2}$ vibration, and this difference is too small to be detected by any but the very best. The other differences vary between these extremes. There are five other records: one for intensity (loudness) discrimination; another for sense of time; another for sense of consonance (harmony); the fifth for tonal memory; and the last for sense of rhythm.

Norms are given for grades five and eight, and for adults.

Scoring the tests.—Keys are provided for scoring each of the tests. The scores are given in terms of the per cent of right responses. For example, in the test

for pitch, if ten errors were made the score would be 90. By means of a table this per cent right score is transferred into a rank score. The rank score represents the standing a person would have on the test among 100 unselected individuals of his group. An adult making a score of 90 on the pitch test would have a rank score of 96; that is, in pitch discrimination there would be only four better than he in a group of 100 unselected adults. If it were an eighth grade child that made a score of 90, his rank would be 98. The same method of scoring and ranking is used with each of the other four tests, except that each has a separate list of ranks to correspond with the various per cent scores. Thus, a per cent score of 90 on the test of intensity corresponds with a rank score of 58; that is, in a group of unselected adults, 42 per cent would surpass the person who obtained a rank score of 58. The keys, table for translating per cent scores into rank scores, as well as a description of the methods of giving and interpreting the score, are given in a Manual of Instruction and Interpretation for Measures of Musical Talent which is furnished by the Columbia Graphophone Company of New York City.

Function of the test.—The Seashore Music Tests were among the first measures of special talent to be constructed. The aim of the test is stated to be not so much the measurement of present attainment as of future possibilities in music. In other words, it is intended to be prognostic. There is some doubt, however, whether they are prognostic of future achievement in music, for validating data are lacking. They do, however, measure six basic capacities which underlie musical talent, particularly performance. It is not at all unlikely that the

Seashore tests are principally valuable in eliminating from future study of music those individuals who are deficient in these six capacities, and who, therefore, would be seriously handicapped from the start.

The tests should, however, not be regarded as absolute criteria; for, as in other fields, a somewhat deficient equipment may be in part compensated for through unusual interest and application.

Although the Seashore tests have been used extensively, only a few studies of their reliability and validity are available. Among the available studies, those of Brennan,³ Brown,⁴ and Stanton⁵ present some interesting results.

In the investigations where the reliability of the separate parts of the test has been examined, the correlations range from + .75 to + .30. These are not so high as the usual correlations for reliability on standard tests; nor are they high enough to be regarded as meeting the criterion of reliability.

Studying the validity of the test presents a more difficult problem, for it is very difficult to get true criteria of musical talent with which to correlate the test scores. Miss Brennan correlated the Seashore scores with the average judgments of four expert musicians on 20 students. The coefficients ranged from + .17 to + .47. Brown got even lower correlations when he used the music instructors' ratings for 90 junior and senior

³ Brennan, Flora M., "The Relation Between Musical Capacity and Performance" (*Psychological Monographs*, 1927, Vol. 36, pp. 190-240).

⁴ Brown, A. W., "The Reliability and Validity of the Seashore Tests of Musical Talent" (*Journal of Applied Psychology*, 1927, Vol. 10, pp. 69-113).

⁵ Stanton, Hazel M., "Eastman School of Music" (*Psychological Studies*, 1929, 1, No. 4).

high school students. Other studies have yielded similar results. One investigation ⁶ gave a higher correlation between intelligence test scores and grades in both musical theory and applied music than similar correlations with the Seashore test.

In contrast with these results are those found by H. M. Stanton at the Eastman School of Music (Rochester, N. Y.). The Seashore tests have been made a part of the entrance requirements to this school as the result of studies conducted by her. Anyone who makes below C (30 per centile on the tests) is not admitted to the school.

In one of these studies the students were grouped into five classes on the basis of the Seashore tests and the Iowa Comprehension Test. More individuals from the upper groups remain in school until they complete their course. Scholarships and honors are almost entirely limited to the upper groups, as is also success in recital work. Fewer and fewer from the lower groups succeed in the work of the school. So few from the worst groups remained to graduate that, as stated, the school since 1928 has refused to admit them.

How can we account for the seeming discrepancy between the results of these studies? First, we should remember that all the studies have shown some relation between test scores and musical talent. Despite all criticisms, the Seashore test does measure musical ability at least to a certain degree. Secondly, at the Eastman School of Music the Seashore tests were combined with the Iowa Comprehension test. In the two

⁶ Highsmith, J. A., "Selecting Musical Talent." (*Journal of Applied Psychology*, 1929, Vol. XIII, pp. 486-493.)

we have a combination of measures of musical talent and general capacity, both of which are essential for success in the Eastman or any other school of music. Therefore, we would expect better results from such a combination than from either type of test used alone.

It is necessary to point out, furthermore, that the coefficients of correlation may be relatively low without at the same time invalidating the practice of excluding those individuals in the lowest levels. It is true that such exclusion may at times work a real injustice against one or several isolated individuals; but for the very large majority the practice is justified.

COURTIS STANDARD SUPERVISORY TESTS IN MUSIC

Description of the tests.—There are two parts to these tests; one part measures ability to recognize characteristic rhythms, and the other tests the recognition of mood from melody. Parts of ten Victor records are played as material for the test, five for each part. A story is first told the class, and then part of a selection is played to illustrate one of the four possible answers to a question in the story. The four possible answers are given on the test blank, and the pupil is to mark the right answer as he interprets it from the music. There are five different sets of questions in the first part dealing with what John or someone else did, and five in the last part describing how John or someone else felt about certain situations.

These are group tests, to be used in grades 4 to 12 inclusive. Blanks are furnished the pupils with the story and questions for use in recording their responses. In

order to provide for retests, two alternative forms are available.

Function of the tests.—These tests are intended to measure musical appreciation, although such appreciation is not often regarded as measurable. Even if the tests should achieve their purpose, it is evident that they deal with only a very limited portion of what is understood to be musical capacity. As such the Courtis tests are suggestive and furnish the materials for an interesting type of instruction in music.

COURTIS STANDARD RESEARCH TESTS

Test 1.—Recognition of Characteristic Rhythms

Rhythm is one of the main elements of music. It has been defined as measured motion. Rhythmic motion also occurs in many of the activities of life. In this test you will be asked to judge from the music played, what life activity is represented.

JOHN'S HOLIDAY

1. It was the first day of the vacation. John had decided to go to a nearby city for a holiday. The music will tell you how John made the journey. Underline the words which tell how the music says he traveled.

1. On foot.	3. On skates.
2. By boat.	4. On horseback.

Follow this story by the Introduction and first eight measures of the Barcarolle—*Victor Record No. 17311*

Test 2.—Recognition of Mood from Melody

Melody is the expression of a thought in music. In this test you will be asked to judge from the music played what John's thoughts were.

5. John's time was now up, so he took his pail and started for home. Listen to the selection and underline the words which best express how the music says John felt when his mother looked at what he had.

1. He was *sorry* he had been cross about going.
2. He was *glad* he had so many berries.
3. He was *ashamed* that he had so few berries.
4. He was *disappointed* that she said nothing.

Follow this story by the melody twice from the beginning of the Serenade Melancholique—*Victor Record No. 6155*

KWALWASSER-RUCH TEST OF MUSICAL ACCOMPLISHMENT

Description of the test.—This test is designed to measure the achievements of pupils in the public school music course of the elementary and high school grades. It includes the following ten parts:

1. Knowledge of musical symbols and terms
2. Recognition of syllable names
3. Detection of pitch errors in a familiar melody
4. Detection of time errors in a familiar melody
5. Recognition of pitch names
6. Knowledge of time signatures
7. Knowledge of key signatures
8. Knowledge of note values
9. Knowledge of rest values
10. Recognition of familiar melodies from notation

The test rests primarily upon the specifications set forth by the Music Supervisors National Conference.⁷

Each pupil receives a booklet and performs the vari-

⁷ Music Supervisors National Conference, Bulletin No. 1, 1921. 64 E. Jackson Blvd., Chicago.

ous tasks designated in the directions of each part. The scoring of the test is entirely objective.

Function of the test.—This is perhaps the best of the “paper and pencil” tests in music, inasmuch as it seeks to conform to the objectives of public school instruction in music; and, further, because as a whole it shows rather high statistical reliability, although some of the parts taken individually are only moderately high in reliability. It is clear, nevertheless, that the test does not go beyond the symbolism of musical notation; which is, in other words, one aspect of the mechanics of musical capacity. Indeed, proficiency in the ten fields listed is no doubt necessary for success in music; but, of course, the functions tested are too restricted in nature to be regarded as thorough tests of musical capacity. However, as analytic measures of accomplishment in *school courses*, they should prove of value.

Drawing

THORNDIKE SCALE FOR THE MERIT OF DRAWING BY PUPILS EIGHT TO FIFTEEN YEARS OLD

(Revised)

Description of the scale.—Thorndike's was the first attempt, in 1913, to measure the quality of free-hand drawing. He collected many specimens of children's drawings and arranged them in order of merit, on the basis of judgments of artists, teachers of art, and students of psychology and education.

During the years 1914–1917, Thorndike obtained ratings for about four thousand specimens from five to

fifteen judges for each. From these four thousand, 303 specimens were chosen for engraving, and for each of these 303 specimens, seventy-five to one hundred additional ratings were obtained. The revised scale of 1923 is based upon these ratings.

Method of using the scale.—The pupils are asked to draw a man, a house, or a snowball fight. Their products may then be compared with the samples of the scale and a rating obtained. "One unit of the scale, i. e., the difference between 2 and 3, or 3 and 4, or 4 and 5, is such a difference in merit as enables seventy-five per cent of artists, teachers of drawing, and students of education to judge that the better drawing is better. Twenty-five per cent will judge wrongly."

Function of the scale.—The scale gives the teacher a series of specimens rated according to merit, the ratings being based upon the judgments of a large majority of individuals qualified to evaluate drawings. It is thus possible to judge the quality of a pupil's drawings with greater precision than would otherwise be the case, as compared with those of other pupils. Unfortunately, there are no age or grade norms, which, if given, would make relative ratings more meaningful.

KLIN-E-CAREY MEASURING SCALE FOR FREE-HAND DRAWING

(Revised)

Description of the scale.⁸—This consists of four scales dealing separately with human figures in action, with rabbits, with houses, and with trees (brush drawings).

⁸ The authors are preparing a second part on design and composition, and a third part on color.

The scales have been extended so that they may be used through the high school. There are 20 samples in the house scale; 19 in the tree scale; 18 in the rabbit scale; and 16 in the boy running scale. Each specimen is accompanied by a legend indicating to the pupil why it is superior to the preceding specimen. The scale, therefore, is designed for use by the pupil himself.

Function of the scale.—Like the Thorndike scale, the Kline-Carey Scale makes available possibilities for more precise ratings of drawings on the basis of quality. It seems that both these drawing scales might be used profitably, one to supplement the other. The Kline-Carey Scale gives no norms.

Conclusions.—In the case of music it will be seen that each of the current measures tests some *portion* of musical talent; that is, a certain quality or qualities which are no doubt essential to success in music. Some, more than others, are significant for the school, inasmuch as they more or less closely conform to school objectives in the subject. If one accepts these objectives as most desirable, then the merit of any single test will depend upon how closely it approaches the objectives. In any event, however, it is very doubtful whether any single scale yet evolved can serve adequately to identify and predict with a high degree of reliability what individuals possess outstanding talent. In spite of their limitations, their usefulness in the class-room and the school system should not be overlooked.

The status of scales of drawing is at present a doubtful one. But even so, they offer a far more sound basis of judgment than individual estimates.

MATERIALS NEEDED

Courtis, S. A., The Courtis Standard Supervision Tests in Music for grades 4 to 12. (S. A. Courtis, 1807 East Grand Blvd., Detroit, Mich.)

Kline-Carey Measuring Scale for Free-Hand Drawing. (Baltimore, The Johns Hopkins Press.) Copies of the four scales and record sheet, 30 cents; booklet containing four scales, directions and record sheet, 60 cents.

Kwalwasser-Ruch Test of Musical Accomplishment. Grades 4 and above. (Bureau of Educational Research and Service, University of Iowa, Iowa City.) Single copy 6 cents; \$5.00 per hundred.

Seashore, C. E., The Measures of Musical Talent (Norms given for 5th and 8th grades and for adults). Six Columbia phonograph records (to be used on any standard machine) with Manual of Directions. Each record \$1.50.

Thorndike Scale for General Merit of Children's Drawings. (Bureau of Publications, Teachers College, Columbia University.) Single copy, 50 cents.

SUPPLEMENTARY LIST OF TESTS

Badger Mechanical Drawing Tests.⁹

Beach Standardized Music Tests.¹⁰

Hillebrand Sight Singing Test.¹¹

Grades 4, 5, and 6.

Hutchinson Music Test; No. 1.⁹

Lewerenz Tests in Fundamental Abilities of Visual Art.¹²

Meier and Seashore Art Judgment Test.¹³

Pressey Technical Vocabularies of the Public School Subjects; Section 15, Music.⁹

⁹ The Public School Publishing Company, Bloomington, Ill.

¹⁰ Bureau of Educational Measurements and Standards. State Teachers College, Emporia, Kansas.

¹¹ World Book Company, Yonkers-on-Hudson, N. Y.

¹² Research Service Co., 7219 Beverly Boulevard, Los Angeles.

¹³ Bureau of Educational Research and Service, University of Iowa.

Spink Grading Chart for Mechanical Drawing¹⁴

Elementary and high school grades.

Torgerson-Fahnstock Music Test.⁹

Grades 4 and above.

SELECTED REFERENCES

Brennan, Flora M., "The Relation Between Musical Capacity and Performance" (*Psychological Monographs*, 1927, Vol. 36, pp. 190-240).

Brown, A. W., "The Reliability and Validity of the Seashore Tests of Musical Talent" (*Journal of Applied Psychology*, 1927, Vol. 10, pp. 69-113).

Dykema, Peter, "Tests and Measurements in Music Education" (*Proceedings of the National Association of Music Teachers*, 1925).

Highsmith, J. A., "Selecting Musical Talent" (*Journal of Applied Psychology*, 1929, Vol. XIII, pp. 486-498).

Kline, L. W., and Carey, G. L., "A Measuring Scale for Free-Hand Drawing, Part I, Representation." *The Johns Hopkins University Studies in Education*, No. 5 (The Johns Hopkins Press); also "A Revision of the Original Scales," No. 5A.

Kwalwasser, J., "The Measurement of the Sense of Rhythm" (*Musical Observer*, June, 1924).

Kwalwasser, J., "Scientific Tests and Measurements Applied to Music" (*Music Supervisor's Journal*, May, 1924).

Kwalwasser, J., "Music" (*Third Yearbook, Department of Superintendence of the National Education Association*, Chapter 14, 1925).

Schoen, M., "Tests of Musical Feeling and Musical Understanding" (*Journal of Comparative Psychology*, Vol. 5, Feb., 1925, pp. 31-52).

Seashore, C. E., "A Survey of Musical Talent in the Public Schools" (*University of Iowa Studies in Child Welfare*, Vol. 1, No. 2, 1920).

¹⁴ Safety Electric Heater Co., 761 Fourth Ave., Faribault, Minn.

Seashore, C. E., "Recent Progress in the Psychology of Musical Talent" (*Proceedings of the Music Teachers' National Conference*, pp. 158-165, 1925).

Seashore, C. E., *The Psychology of Musical Talent* (New York, Silver, Burdette and Company, 1919).

Stanton, Hazel M., "Eastman School of Music" (*Psychological Studies*, 1929, Vol. 1, No. 4).

Thorndike, E. L., "A Scale for General Merit of Children's Drawings" (*Teachers College Bulletin*, Fifteenth Series, No. 6, December, 1923).

Thorndike, E. L., "The Measurement of Achievement in Drawing" (*Teachers College Record*, Vol. 14, Nov., 1915, pp. 345-382).

Trabue, M. R., "Scales for Measuring Judgment of Orchestral Music" (*Journal of Educational Psychology*, Vol. 14, Dec., 1923, pp. 545-561).

Weaver, A. T., "Experimental Studies in Vocal Expression" (*Journal of Applied Psychology*, Vol. 8, pp. 23-51, 159-186, 1924).

CHAPTER XIII

SECONDARY SCHOOL MATHEMATICS

The problem of measurement in mathematics.— Measuring results of the teaching of secondary school mathematics is complicated by the fact that the objectives of the subject are neither universal nor well defined. The place and importance of algebra and geometry are not agreed upon. Shall mechanics or reasoning be emphasized? If the mechanics are to be stressed, then what shall be the relative importance of the various phases? The difference of opinion with respect to these and other relevant questions in high school mathematics is reflected in the diversity found in courses of study and in textbooks. It is very probable, however, that the 1923 report of the National Committee on Mathematical Requirements¹ will exert some influence in stabilizing and clarifying content and method of secondary school mathematics.

The tests of secondary school algebra and geometry, developed for the most part during the last dozen years, have accepted conditions pretty much as they found them and have proceeded to measure principally the fundamentals and the mechanics. As in history and geography, measurement of the broader aspects of the subject, such

¹ "The Reorganization of Mathematics in Secondary Education." *National Committee on Mathematical Requirements*, Auspices of the Mathematical Association of America. Published by the Association, 1923.

as the effect of mathematical training on general situations and problems, has not been attempted. But perhaps it is not the province of mathematical tests to do so. If the influence of a specific subject on one's general attitude and behavior is commensurate with his grasp of the subject, then it seems the important thing is the measurement of that subject. It is very doubtful whether achievement tests should even attempt to measure transfer values.

General

ROGERS TEST OF MATHEMATICAL ABILITY

Description of the test.—This test is designed to measure the "mathematical intelligence" of pupils who have had five months of formal algebra and no formal geometry. It is to be used in the first year of senior high school or the third year of junior high school as a measure of probable success in more advanced courses in mathematics. There are six parts to the test: (1) a geometry test; (2) algebraic computation test (test 1 and 2); (3) interpolation test (test 1 and 2); (4) superposition test (test 1 and 2); (5) Trabue language scales (L and J); and (6) mixed relations test.

Method of using the test.—Before each test is given, a careful explanation of the processes involved in the problems of the test is made by the tester. The explanations for each test are outlined in the manual of directions. Eight minutes are allowed for explanation in the geometry test and twenty-two minutes for the test itself. This test contains six problems. In each problem

certain statements are given, and from these the pupils are to answer a question and then state the reason or give the proof. All the reasons are to be chosen from a series of facts presented on the pages opposite the problems.

One-fourth of a minute is given for explanation of the algebraic computation tests; three minutes are allowed for the first test and seven minutes for the second. There are eleven simple problems in the first test and seven more difficult problems in test 2.

A similar method of distributing the time between explanation and work on the test problems is used in each of the other tests. The interpolation test consists in supplying certain missing figures in a number series. For example, series are given such as the following:

A.	...	1	3	5	7	..	11	13	15	17	..	21
D.	...	1	8	15	..	29	36	43	..	57	64	71
R.	...	11	66	121

in which the missing numbers are to be filled in to complete the series. The superposition test consists in locating the position of a circle in the corner of a given parallelogram by revolving a similar figure in imagination to fit on a given base line. The language test contains two parts of the Trabue language completion test. The mixed relations test is a statement of a proportion. The first two terms consist of words related in some way, with a fourth term to be filled in to bear the same relation to the third as the first does to the second.

Function of the test.—“The Rogers test represents six measuring rods of abilities, which after an intensive study of the activities demanded by high school mathe-

matics were selected as possessing the highest predictive power.”² The tests have been used principally to advise pupils regarding further study of high school mathematics and to section mathematics classes on the basis of ability. For both these purposes, it would be well to use the Rogers tests together with a suitable test of intelligence.

KELLEY MATHEMATICAL VALUES TEST ALPHA

By means of one questionnaire to teachers of mathematics and another to a group of “capable and successful Americans,” T. L. Kelley investigated the value to be derived from the study of algebra.³ He then constructed a test consisting of thirty-eight problems to measure the thirteen fundamental mathematical values most often cited. Mathematical Values Test Alpha is partly a test in general mathematics and partly in algebra. It differs greatly from traditional tests in that the problems cover a broader field than the usual examination in algebra. Although the test may not have much value as a measure of achievement, it might help to estimate the “broader” values and interests to be derived from the study of high school mathematics.

Algebra

HOTZ FIRST YEAR ALGEBRA SCALES

Description of the scales.—These scales, devised by H. G. Hotz of the University of Arkansas, consist of a

² “The Rogers Test of Mathematical Ability, Manual of Directions” (Teachers College, Columbia University, 1921).

³ Kelley, Truman L., “Values in High School Algebra and Their Measurement” (*Teachers College Record*, Vol. XXI, No. 3, May, 1920).

series of five sets of exercises. The first set is made up of exercises in addition and subtraction, the second of exercises in multiplication and division, the third in equations and formulas, the fourth of graphs and the fifth of problems. The first two sets are designed "to test the achievement of students in the fundamental operations involving integral, fractional, and radical expressions; the second two, to test the ability of students in handling the instruments of quantitative thinking; while the last is composed of verbal problems of the type usually stressed in first year algebra."⁴ The exercises in each set are arranged in order of difficulty. The first problem of each scale is so easy that it can be solved by practically every pupil in the class. But each succeeding problem is increasingly difficult, so that the last few in each scale will be solved by a very small percentage of students who attempt them.

Method of using the scales.—There are two series of all scales: Series A and Series B. Series B, which is the longer of the two, contains from eleven to twenty-five problems in each test, while Series A, about half as long, has from eight to twelve problems in each test. The latter covers the same range of difficulty as the longer set; and it is recommended that A be used if all five scales are to be given. If time is limited, Hotz recommends that the equation and formula scale be selected, for it is the most comprehensive of the group. Series B is suggested for making an analysis of the difficulties of a class or individual. It is further recommended that if the whole scale is to be given, the tests should be used

⁴ Hotz, H. G., "Teachers' Manual for First Year Algebra Scale" (*Teachers College Publication, Columbia University*).

in rotation somewhat as follows: at the end of three months, addition and subtraction, equations and formulas; at the end of six months, multiplication and division problems; at the end of nine months, equations and formulas (repeated), graphs. The time allowances are twenty minutes for each of the first three sets of exercises and twenty-five minutes each for the last two in Series A. In Series B forty minutes are allowed for each exercise except for graphs for which the time limit is twenty-five minutes.

The scoring is made simple by neglecting entirely the principles of solution and by scoring for correct answers only. Answers are provided in the manual.

Function of the scales.—The Hotz Scales have many characteristics of a good test. The subject-matter is based upon a careful analysis of materials being taught; the materials were standardized by obtaining results over a wide distribution of schools and pupils; and the problems are scaled in difficulty. Furthermore, the time limits are such as to make these scales measures of power in algebra, rather than of speed, which is entirely secondary.⁵ The scales have also been shown to possess rather high reliability. There are perhaps two principal criticisms of these scales: First, there are no alternative forms; the value of a test is always enhanced when there are such. Second, the scoring of correct answers only frequently disregards the mastery of correct principles. Yet the scoring of these scales may

⁵ The tests devised by Rugg and Clark may be used if a measure of speed is desired. A *power* test indicates how difficult a type of problem the pupil is capable of solving. The *speed* test shows a pupil's skill at a given level of difficulty, since all the problems are of the same or approximately equivalent difficulty.

be justified, for otherwise it is extremely difficult to meet the demands of objectivity.

TENTATIVE MEDIAN STANDARDS OF ACHIEVEMENT

Series A

	3 mos.	6 mos.	9 mos.
Addition & Subtraction.....	5.0	6.8	7.9
Multiplication & Division....	5.3	6.3	7.9
Equation & Formula.....	4.9	7.1	7.8
Problems	4.3	4.9	5.6
Graphs	2.8 (4½ mos.)		5.6

Series B

	3 mos.	6 mos.	9 mos.
Addition & Subtraction.....	9.7	12.9	14.4
Multiplication & Division....	9.6	14.0	16.3
Equation & Formula.....	7.8	14.3	16.0
Problems	5.4	6.5	7.5
Graphs	3.7 (4½ mos.)		7.2

DOUGLASS STANDARD DIAGNOSTIC TESTS FOR
ELEMENTARY ALGEBRA

Description of the tests.—These tests, devised by H. R. Douglass, are designed to measure the fundamental operations of elementary algebra. As the result of a questionnaire sent to teachers of mathematics in high schools, normal schools, and colleges throughout the country, ten exercises were constructed to cover the widest possible range of abilities in each of the four operations in algebra ranked as most fundamental by these teachers. These four fundamental operations, incorporated in Series A, are:

1. Collection of terms (addition and subtraction)
2. Multiplication

3. Division
4. Solution of simple equations

In order to get as wide a distribution of types of problems as possible, fifteen standard texts in algebra were studied. The problems were selected from the different texts; and, to avoid any effect of practice, each exercise was slightly changed from its original form.

Series B is designed to measure progress in the following:

1. Fractions
2. Factoring
3. Formulas and fractional equations
4. Simultaneous equations
5. Graphical representation and interpretation
6. Square roots, exponents, and radicals
7. Quadratic equations

The selection of this material was based upon the content of modern textbooks, the suggestions of some twenty teachers of mathematics in large high schools, and upon the recommendations of the National Committee on Mathematical Requirements.

Method of using the tests.—Series A should be given near the end of the first term of instruction in algebra, and Series B near the end of the second term of instruction.

Function of the tests.—As the name implies these are diagnostic tests. Though there are stated time limits, they are such as to make speed a negligible factor, so that the tests measure power primarily. This must be true if the tests are to fulfil their diagnostic purpose. By using all parts of the tests, it is possible to enumerate the specific errors and their frequencies, which will in-

dicate the direction to be taken by remedial teaching. Such analysis will show errors in exponents, signs, operations, coefficients, etc.⁶ These scales show a fairly high degree of reliability, though not so high as the Hotz Scales. Both, however, have many characteristics of a good standard test.

TENTATIVE NORMS FOR THE DOUGLASS TESTS

<i>Test Number . . .</i>	1	2	3	4	5	6	7
Series A	7.8	7.1	6.5	7.3			
Series B	2.4	4.1	3.1	3.6	2.5	2.7	3.4

ILLINOIS STANDARDIZED ALGEBRA TESTS

Description of the tests.—There are four different tests in this series. In all the exercises, the pupil is required to solve for “*x*.” The first consists of twenty simple equations in addition and subtraction; the equations of test II call for a knowledge of transposing, of test III for the removal of the parentheses, and of test IV for the reduction of fractions. Four minutes are allowed for the first test, five minutes for the second, nine minutes for the third, and ten minutes for the fourth.

	FIRST SEMESTER		SECOND SEMESTER		THIRD SEMESTER	
	No.		No.		No.	
	problems	No.	problems	No.	problems	No.
Test I.....	9.8	5.0	10.5	6.4	11.7	8.7
“ II	10.8	4.6	11.8	6.4	12.6	8.3
“ III.....	11.0	3.6	12.2	5.5	13.9	7.3
“ IV.....	8.8	1.0	11.3	3.8	13.2	5.7

Function of the tests.—The subject-matter of these tests is based upon a study of the processes most used

⁶ For frequencies, see Ruch and Stoddard, *op. cit.*, p. 79.

EXERCISES FROM EACH OF THE FOUR TESTS OF THE
ILLINOIS STANDARDIZED ALGEBRA TESTS

TEST I

1. $5x - 7 = 3x - 15$

2. $-7x + 15 = 5x - 57$

3. $13x + 16 = -9x - 19$

4. $17x - 23 = -11x + 65$

5. $8x - 9 = 11x - 3$

TEST III

1. $-4(11x - 7) = 33x - 126$

2. $3(-3x + 4) = 7x - 100$

3. $-7(3x + 9) = -6x - 13$

4. $9(7x - 19) = -42x + 354$

5. $-7(3x - 5) = 14x - 7$

TEST II

1. $13x - 6x = 70 - 14$

2. $-11x + 7x = 45 - 25$

3. $9x + 4x = -41 - 30$

4. $13x - 7x = -23 + 17$

5. $5x - 17x = 35 - 83$

TEST IV

1. $\frac{3x - 14}{4} = \frac{8x - 15}{6}$

2. $\frac{-(-5x + 3)}{4} = \frac{8x - 7}{3}$

3. $\frac{-(3x + 5)}{7} = \frac{-(-5x - 3)}{4}$

4. $\frac{5x - 4}{7} = \frac{-(-3x + 9)}{5}$

5. $\frac{11x - 4}{7} = \frac{13x - 3}{8}$

by pupils in solving algebraic exercises. Clearly enough, the scope of these tests is restricted; and because of their limited range, their value is also limited. The reliability of the tests is rather high, thus indicating that the processes they do measure are measured rather consistently. This fact alone, however, does not make the Illinois test so valuable as those of Hotz and Douglass for purposes of analysis and remedial teaching.

Geometry

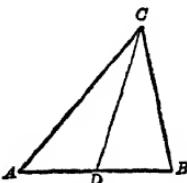
MINNICK GEOMETRY TESTS

Description of the tests.—Four tests, A, B, C, D, compose the series. Test A involves construction; five propositions are stated, for each one of which the pupil is to construct a figure. In the four exercises of test B, the figure is presented and the theorem given. The pupil must state what is given and what is to be proved. In test C, which has four items, a figure is given, certain facts about it are presented, and the pupil is required to give as many more facts about it as he can. In test D, having three exercises, figures are given, facts about them are presented, and the hypotheses are stated. The pupil is required to state the proof.

Method of using the tests.—The method of marking these tests yields two separate scores. One, the number of facts given correctly, is called the positive score. The other, the number of incorrect or unnecessary statements, is the negative score. The latter is intended to serve the purpose of diagnosis and to give the teacher the necessary information with regard to types of errors and special difficulties. Standards are presented for both

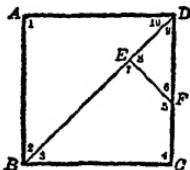
L. Draw the figure for the following proposition:

If two radii of a circle are perpendicular, and a tangent to the circle cuts these radii produced at points A and B, the other tangents drawn from A and B are parallel.



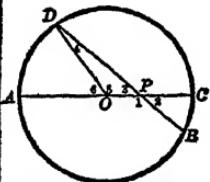
Given:

To Prove:



IV. State what is given and what is to be proved in the following proposition:

An angle of a triangle is a right angle, an acute angle, or an obtuse angle, according as the median drawn from the vertex of the angle is equal to, greater than, or less than one-half of the opposite side.



II.

GIVEN: The square ABCD, the diagonal BD,
 $EB = CD$ and EF is perpendicular to BD.

State as many more facts about this figure as you can.

II.

GIVEN: P is any point within the circle O,
AC is a diameter through P,
BD is any other chord through P, OD is a radius.

To prove that $AP > DP$.

Other known facts:

$$\angle 6 = \angle 3 + \angle 4.$$

$$PD - OD < OP.$$

$$OD + OP > DP.$$

$$\angle 3 = \angle 4.$$

$$AO = OD.$$

$$\angle 5 + \angle 6 = 180^\circ$$

$$AO + OP = AP.$$

Proof:

scores. The scoring method is cumbersome and also rather subjective. For example, in one instance three judges differed by 22 points in the positive score and by 11 points in the negative.⁷

Function of the tests.—These tests are intended to measure only the formal aspects of high school plane geometry. This they do, but perhaps to an insufficient degree. They are in part diagnostic; but the scoring made necessary to facilitate diagnosis seriously weakens their objectivity. The reliability of these tests is moderately good.

SCHORLING-SANFORD ACHIEVEMENT TEST IN PLANE GEOMETRY

Description of the test.—The items of this test were originally made up by R. Schorling in 1921. Subsequently, three revisions were made by V. Sanford, the final revision being made more nearly in accord with the recommendations of the National Committee on Mathematical Requirements.

There are two equivalent forms, A and B, each having five parts with twelve questions in each. The five parts are as follows:

1. Completing sentences (factual)
2. Drawing conclusions from given data
3. Judging the correctness of conclusions
4. Analyzing constructions
5. Computations (angles, areas, lengths, etc.)

Fore-exercises are provided for all parts.

⁷ Morrison, J. C., "The Use of Standard Tests and Scales in the Plattsburgh High School" (*University of the State of New York Bulletin*, No. 784, 1923).

Method of using the test.—The time limit of each form is fifty-two minutes, divided into definite time limits for each part. The fore-exercise requires about ten minutes, so that it will be desirable to allow two class periods for the test. The scoring is objective and may be done rapidly. Norms are provided for both forms.

Function of the tests.—Unlike the Minnick tests, the Schorling-Sanford tests are not intended for diagnostic purposes. They are to be used as a final achievement examination in plane geometry, the material being drawn from all five books of plane geometry.

Though these tests are not intended for diagnosis, they can nevertheless contribute toward the improvement of teaching technique, for an analysis and tabulation of errors by the teacher will be of considerable value in that respect.

The reliability of the Schorling-Sanford tests is moderately high.

COLUMBIA RESEARCH BUREAU PLANE GEOMETRY TEST

Description of the test.—This examination contains two types of questions. Part I is of the "true-false" kind, made up of 65 statements concerning such matters as propositions, corollaries, loci, formulas, constructions, etc. Part II consists of 35 problems, ranging from the very easy to the very difficult. "The student indicates his understanding and solution of the problems by means of numerical answers; but no burden of an arithmetical or computational sort is put upon the candidate, since he is allowed to indicate the necessary

operations by the use of formulas and equations. It is a test of geometrical reasoning and not of exactness in arithmetical calculations.' Both parts include materials from every part of plane geometry and are thus believed to give a reliable measure of reasoning ability with the materials of the subject.

Method of using the test.—The working time of the test is 60 minutes—20 minutes for part I and 40 minutes for part II. Each student receives a booklet, starts at a given signal and continues on one part at a time without interruption. The scoring of the test is easy and objective.

For the purpose of supplementing the results of the regular Columbia Research Bureau test, there is the so-called "Augmented Test," having four parts: (1) loci; (2) converses; (3) definitions; and (4) demonstrations. These are not intended as alternative forms; they are designed solely for use in conjunction with the regular forms A and B, when a check and supplementary information are desired.

Function of the test.—Forms A and B are intended principally as a final examination in high school courses in plane geometry. According to the authors, they may also be used by colleges as an aid in selecting from among candidates those who have had an adequate preparation in plane geometry, as well as in the identification of the especially gifted or deficient in this subject. The authors also state that the test serves the purposes ordinarily served by the good standardized test: namely, (1) for assigning marks; (2) for establishing comparable norms; (3) for educational and vocational counseling; (4) for educational research.

The reliability of the test was studied with 1,349 pupils. The results show part I to have rather low reliability, but part II to have high reliability. The reliability of the combined scores of I and II, however, is very high. The relatively low reliability of part I may be due in part to the fact that the items are of the "true-false" type, where uncertainty and consequent guessing might well make for inconsistency of results.

Conclusions.—It is possible that some, among them teachers of mathematics, will regard the tests of algebra and geometry as unsatisfactory, inasmuch as their scope seems to be limited. It might be said that the tests do not measure important habits and skills derived from the study of mathematics. But these derived benefits are at present rather intangible and indefinite, so that at this time we must rest content with the measurement of the specific skills which are being taught. Yet, even in this, difficulties are encountered, for there are differences of opinion with respect to relative importance of various phases and the emphasis they should receive. The author of a test must, therefore, be guided by what seems to be the predominant practice, as indicated by teachers and as found in textbooks.

MATERIALS NEEDED

Columbia Research Bureau Plane Geometry Test. (Yonkers-on-Hudson, N. Y., World Book Company.) Package of 25 examinations, with key and manual of directions, \$1.20, either form. Specimen set, 25 cents.

Douglass Standard Diagnostic Tests for Elementary Algebra. University of Oregon. Series A, \$1.60 per hundred; Series B, \$3.50 per hundred, including key and class record sheet.

Hotz Algebra Scales. Bureau of Publications, Teachers College, Columbia University. Each scale 70 cents per hundred, except graph scale which is \$1.25 per hundred. Manual of direction for examiner, 75 cents.

Illinois Standardized Algebra Tests. (Bloomington, Ill., The Public School Publishing Company.) \$2.50 per hundred; specimen set, 15 cents.

Kelley Mathematical Values Test, Bureau of Publications, Teachers College, Columbia University. One set of scales, 40 cents; test blanks, 5 cents each.

Minnick Geometry Tests. (Bloomington, Ill., The Public School Publishing Company.) \$2.50 per hundred; specimen set, 20 cents.

Rogers Test for Diagnosing Mathematical Ability. Bureau of Publications, Teachers College, Columbia University. Manual of directions and stencils, 50 cents; test booklets, \$7.00 per hundred; specimen set, 10 cents.

Schorling-Sanford Test in Plane Geometry. Bureau of Publications, Teachers College, Columbia University. Directions and stencils, 50 cents; test booklets, \$7.00 per hundred; specimen set, 10 cents.

SUPPLEMENTARY LIST OF TESTS

American Council Solid Geometry Test.⁸

American Council Trigonometry Test.⁸

Columbia Research Bureau Algebra Test.⁸

Hart Diagnostic Tests and Drills in First Course Algebra.¹¹

Hart Geometry Tests.¹¹

Iowa Placement Examinations, Revised, Mathematics.⁹

McMindes Plane Geometry Tests.⁹

Orleans Algebra Prognosis Test.⁸

Orleans Geometry Prognosis Test.⁸

⁸ World Book Company, Yonkers-on-Hudson, N. Y.

⁹ The Public School Publishing Company, Bloomington, Ill.

¹⁰ Bureau of Administrative Research, College of Education, University of Cincinnati.

¹¹ D. C. Heath & Co., Boston.

Renfrow Diagnostic Tests in Plane Geometry.¹⁰
Schorling-Clark-Lindell Instructional Tests in Algebra with
Goals for Pupils of Varying Abilities.⁸
Seattle Solid Geometry Tests.⁹
Webb Geometry Test.⁹

SELECTED REFERENCES

Douglass, H. R., "The Derivation and Standardization of a Series of Diagnostic Tests for the Fundamentals of First Year Algebra" (*University of Oregon Publication*, Vol. 1, No. 8, 1921).

Douglass, H. R., "The Douglass Standard Diagnostic Tests for Measuring Achievement in First-Year Algebra—Revisions and Extensions" (*University of Oregon Publication*, Vol. 2 No. 5, University of Oregon, 1924).

Eells, W. C., "Hotz Algebra Scales in the Pacific Northwest" (*The Mathematics Teacher*, Vol. 18, Nov., 1925, pp. 418-427).

Harris, E., and Breed, F. S., "Comparative Validity of the Hotz Scales and the Rugg-Clark Tests in Algebra" (*Journal of Educational Research*, Vol. 6, Dec., 1922, pp. 393-411).

Hotz, H. G., *First Year Algebra Scales* (Contributions to Education, No. 90, Teachers College, Columbia University, 1918).

Kelley, T. L., "Values in High School Algebra and Their Measurement" (*Teachers College Record*, Vol. 21, No. 3, pp. 246-290, 1920).

Minnick, J. H., "An Investigation of Certain Abilities Fundamental to the Study of Geometry" (*University of Pennsylvania*, 1918).

National Committee on Mathematical Requirements. "The Reorganization of Mathematics in Secondary Education," 1923.

Rogers, A. L., "Experimental Tests of Mathematical Ability and Their Prognostic Value" (Bureau of Publications, Teachers College, Columbia University, 1918).

Rugg, H. O., and Clark, J. R., "Scientific Method in the Reconstruction of Ninth Grade Mathematics" (*Supplementary Educational Monographs*, Vol. 2, No. 1, 1918).

Sanford, V., "A New Type Final Geometry Examination" (*The Mathematics Teacher*, Vol. 18, January, 1925, pp. 22-36).

Schorling, R., and Clark, J. R., "A Program of Investigation and Coöperative Experimentation in the Mathematics of the Seventh, Eighth and Ninth School Years" (*The Mathematics Teacher*, Vol. 14, May, 1921, pp. 264-275).

Thorndike, E. L., "The Nature of Algebraic Abilities" (*The Mathematics Teacher*, Vol. 15, January, 1922, pp. 6-15; Feb., 1922, pp. 79-92).

Williams, L. W., "Illinois Standardized Algebra Test" (*Journal of Educational Research*, Vol. 3, January, 1921, pp. 75-76).

CHAPTER XIV

SECONDARY SCHOOL SCIENCE

Problems in the measurement of science.—In spite of the fact that the construction of tests in science appear at first glance to be a relatively easy matter, they present difficulties no less marked than in history or mathematics. There is in science the usual lack of agreement with respect to the content and the organization of units in the several subjects.¹ The conflict between emphasis upon information, on the one hand, and upon reasoning and the development of “scientific habits and attitudes” on the other, will lead some to adverse criticism of any current standard tests, for these habits and attitudes are quite intangible at present and scarcely touched upon in the tests. The measurement of specific information, however, is more easily accomplished; yet here, as in other subjects, the complete validity of the tests must wait on the more nearly definite and universal statement of aims and objectives; and this statement must come primarily from those responsible in the field of science. Among the current tests, many authors have distributed their items so as to include the specific, recognized skills, such as numerical computa-

¹ See “Reorganization of Science in the Secondary Schools,” *U. S. Bureau of Education, Bulletin 36* (1920). Also “The Reorganization of High School Science,” by Barber, F. D., (*School Science and Mathematics*, Vol. 23, 1923, pp. 247-262). Also “The Problem of Science Teaching in the Secondary Schools,” by Millikan, R. A., (*School Science and Mathematics*, Vol. 25, 1925, pp. 966-975).

tion, laboratory exercises, informational data, and reasoning with the data.

General Science

VAN WAGENEN READING SCALES—GENERAL SCIENCE, SCALES A AND B

Description of the scales.—The form of these General Science Scales is similar to the other Van Wagenen scales already described in the chapters on History and Geography. There are fifteen short paragraphs on various topics in general science, followed by sets of four to six questions based on the subject-matter of the paragraphs. Scale A is weighted somewhat with biology, while Scale B leans to chemistry and physics. The paragraphs are divided into three groups of five each. The pupil is directed to read each paragraph carefully. “Then read the statements below it and put a check mark (✓) on the dotted line in front of each statement which contains an idea that is in the paragraph or that can be derived from it.”

Scoring the scales.—The Van Wagenen Scales all employ a complicated scoring system involving weightings. The number of errors is counted separately for each of the three groups of paragraphs. The uncorrected score for a pupil is obtained from the key on the basis of the number of errors made in the questions to the paragraphs of group III. This first score is corrected in relation to the number of errors made in the answers to the questions in group II and recorrected likewise for errors in group I to give the final score. Tentative norms are given for grades 8 to 12 inclusive.

The advantages of a weighting system are extremely doubtful, for scoring by such a method is a long process and does not yield results superior to those arrived at through a simple method.² In fact, relatively few tests employ weighted scores.

Function of the scales.—This type of scale requires the pupil to abstract the meaning of the material, inasmuch as some of the questions refer to information contained within the paragraph, while others refer to certain implications of the paragraph. If the subject-matter were entirely new, the scales would then measure ability to comprehend reading materials of a scientific nature. Since some of the material is likely to be more or less familiar to the high school student, the score may be in part a measure of his stock of information. The scales are, however, principally measures of ability to comprehend scientific reading matter.

DOWNING RANGE OF INFORMATION TEST IN SCIENCE

Description of the test.—The test devised by E. R. Downing consists of a list of fifty words or phrases selected from the various sciences. The terms are well distributed between physiology, geography, biology, physics, and chemistry. The words are arranged alphabetically on the test sheet. The pupil is directed to put an "E" beside the words and phrases that he can explain or define, an "F" beside the ones he has heard or read about, the meaning of which is not clear, and an

² See "Is It Necessary to Weight Exercises in Standard Tests?" by Douglass, H. R. and Spencer, P. L., (*Journal of Educational Psychology*, Vol. 14, Feb., 1923, pp. 109-112).

“N” beside those that are new. He is then directed to explain or define the first five marked with an “E.” The pupil is given all the time needed to mark every term.

Scoring the test.—The answers are scored on the basis of the number of words marked in each of the three groups “E,” “F,” and “N,” except that the “E” list is reduced by the per cent of words wrongly defined, and the reduction is added to the “F” list. For example, if a pupil has 15 words marked with an “E,” 15 with an “F,” and 10 with “N,” and one of his definitions is wrong, his score should be E-12, F-18, N-10.

SECTION FROM THE REVISED RANGE OF INFORMATION
TEST IN SCIENCE

By Dr. Elliot R. Downing

Please put an E beside words and phrases (on the list below) that you can explain or define, an F beside those you have heard or read about, the meaning of which is not clear, and an N beside those that are new. Explain or define the first five you mark with an E, on the back of this sheet.

	<i>Mark</i> <i>Here</i>		<i>Mark</i> <i>Here</i>
No.		No.	
1	Adaptation	26	Inoculation
2	Atom	27	Instinct
3	Buoyancy	28	Law of gravitation
4	Candle power	29	Law of the lever
5	Center of gravity	30	Law of the pulley

Function of the test.—The Downing test provides a ready device for a rapid survey of general information in general science. The list is, of course, but a sampling; its validity as a measure of information, therefore, depends upon how representative and important the selection of terms is.

The advantages of a weighting system are extremely doubtful, for scoring by such a method is a long process and does not yield results superior to those arrived at through a simple method.² In fact, relatively few tests employ weighted scores.

Function of the scales.—This type of scale requires the pupil to abstract the meaning of the material, inasmuch as some of the questions refer to information contained within the paragraph, while others refer to certain implications of the paragraph. If the subject-matter were entirely new, the scales would then measure ability to comprehend reading materials of a scientific nature. Since some of the material is likely to be more or less familiar to the high school student, the score may be in part a measure of his stock of information. The scales are, however, principally measures of ability to comprehend scientific reading matter.

DOWNING RANGE OF INFORMATION TEST IN SCIENCE

Description of the test.—The test devised by E. R. Downing consists of a list of fifty words or phrases selected from the various sciences. The terms are well distributed between physiology, geography, biology, physics, and chemistry. The words are arranged alphabetically on the test sheet. The pupil is directed to put an "E" beside the words and phrases that he can explain or define, an "F" beside the ones he has heard or read about, the meaning of which is not clear, and an

² See "Is It Necessary to Weight Exercises in Standard Tests?" by Douglass, H. R. and Spencer, P. L., (*Journal of Educational Psychology*, Vol. 14, Feb., 1923, pp. 109-112).

“N” beside those that are new. He is then directed to explain or define the first five marked with an “E.” The pupil is given all the time needed to mark every term.

Scoring the test.—The answers are scored on the basis of the number of words marked in each of the three groups “E,” “F,” and “N,” except that the “E” list is reduced by the per cent of words wrongly defined, and the reduction is added to the “F” list. For example, if a pupil has 15 words marked with an “E,” 15 with an “F,” and 10 with “N,” and one of his definitions is wrong, his score should be E-12, F-18, N-10.

SECTION FROM THE REVISED RANGE OF INFORMATION
TEST IN SCIENCE

By Dr. Elliot R. Downing

Please put an E beside words and phrases (on the list below) that you can explain or define, an F beside those you have heard or read about, the meaning of which is not clear, and an N beside those that are new. Explain or define the first five you mark with an E, on the back of this sheet.

<i>Mark</i> No. <i>Here</i>	<i>Mark</i> No. <i>Here</i>
1 Adaptation	26 Inoculation
2 Atom	27 Instinct
3 Buoyancy	28 Law of gravitation
4 Candle power	29 Law of the lever
5 Center of gravity	30 Law of the pulley

Function of the test.—The Downing test provides a ready device for a rapid survey of general information in general science. The list is, of course, but a sampling; its validity as a measure of information, therefore, depends upon how representative and important the selection of terms is.

GRIER RANGE OF INFORMATION TEST

This is very similar to the Downing test, except that there are three lists of one hundred terms each. One list is devoted to physiology, a second to zoölogy, and the third to botany.

RUCH-POOPENOE GENERAL SCIENCE TEST

Description of the test.—This test, primarily of eighth and ninth grade general science, is available in two forms, A and B, each of which consists of two parts. The first part deals with information in chemistry, physics, zoölogy, botany, astronomy, geology, physiography, and physiology. The questions, fifty in number, are of the multiple-choice type, with seven responses to choose from. The second part of the test has twenty drawings and diagrams, with eighty statements concerning them, these statements being the completion type; that is, in each statement one or more critical words are omitted. The pupil is required to supply these words, and his ability to supply the correct information is the measure of his comprehension of the situation.

The materials of these tests are based upon twenty-three textbooks and manuals; and—in part II only—upon the ratings of experienced teachers in general science.

Scoring the tests.—The tests are easily administered, and the total working time is forty minutes. The scoring is simple and objective; in part I the score is the total number of correct responses; in part II it is the number correct divided by 2. Percentile norms are provided for January and June.

		Ruch-Popenoe	Form A
24	The length of a meter in inches is about	12 18 24 2 39 47 144	24
25	A general term for any living thing is a	plant cell larva animal organism mammal nucleus	25
26	A violent circular windstorm of small area is a	cyclone tornado monsoon trade wind norther blizzard equinox	26
27	A collection of similar cells is called an	organism tissue organ gland muscle sense-organ function	27
28	The watt is the unit of measurement of	resistance current velocity power potential inductance friction	28
29	The unborn young of an animal is termed the	larva embryo pupa adult chrysalis ovum sperm	29
30	An example of a chemical element is	water glass mercury carbon dioxide ammonia nitric acid sugar	30

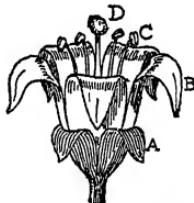


FIGURE 2

In this diagram of a typical flower

- a The petals (the corolla) are marked by the letter a
- b The stamens are marked by the letter b
- c The sepals (the calyx) are marked by the letter c
- d The pistil is marked by the letter d



FIGURE 3

- a The North Star (Polaris) is marked by the letter a
- b The Big Dipper is marked by the letter b

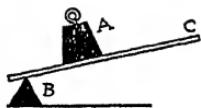


FIGURE 4

- a In this lever the power or force is applied at a
- b The fulcrum is placed at the point marked b
- c The mechanical advantage of a lever of this class is always than 1. c

SAMPLE EXERCISES FROM THE RUCH-POOPENOE GENERAL SCIENCE TEST

Function of the tests.—These tests are intended to measure the subject-matter of the usual first year course in general science. The sampling appears to be rather

wide. The authors of the tests suggest that they may be used in connection with the usual school problems such as in the determination of promotions, for purposes of classification, for assigning marks, for comparative purposes, etc. The tests show a fairly high degree of reliability.

SUBJECT-MATTER ANALYSIS OF THE RUCH-POPOE GENERAL SCIENCE TEST IN PERCENTAGES

Biological science (botany, physiology and zoölogy)	30%
Chemistry	12%
Physics and mechanical applications.....	38%
Earth science (agriculture, astronomy, geology, and physiography)	20%

DVORAK GENERAL SCIENCE TESTS

This test is intended for much the same purpose as the Ruch-Popenoe tests, with one added feature: scale R-1 may be used for diagnostic purposes early in the year. There are three forms, each having sixty questions of the multiple-choice type. These questions are also based upon an analysis of textbooks. It appears that the materials were standardized with care, so that the reliability of the tests is rather high.³ The Dvorak tests do not involve speed of work, for each pupil is allowed as much time as he needs. Detailed percentile norms are provided.

It will be seen, of course, that both the Dvorak and

³ Curtis, F. D., in *Some Values Derived from Extensive Reading of General Science*, commends the Dvorak tests because of "their careful standardization, ease of administering and scoring, and all-round excellence," Contributions to Education, No. 163, Teachers College, Columbia University, 1924.

the Ruch-Popenoe tests are much more comprehensive and searching than the Downing or the Grier tests.

Chemistry

POWERS GENERAL CHEMISTRY TEST

Description of the test.—The test, of which there are two forms, A and B, is designed to measure the achievement of pupils in high school chemistry. The items were chosen from high school textbooks and subjected to experiment for four years, after which time 134 items were retained from the original 350. Each form, of two parts, contains 67 items arranged in the order of increasing difficulty. Part I, composed of 30 items, tests range of information, including biography, chemical properties, chemical composition, commercial processes, and terminology. Part II has 37 items testing ability to write formulas and equations, to give chemical names of substances, and to perform simple calculations.

Scoring the test.—The working time for each form of the test is 35 minutes. This time limit is sufficient to make it a measure of power, for it has been shown that increasing the time allowance does not appreciably increase the score. The student's score, easily and objectively found, is the number of items answered correctly. Percentile norms are provided in the manual of directions.

Function of the test.—The Powers Chemistry Test is at present regarded by many as one of the best in the field of science and may be used to marked advan-

tage for any of the purposes to which standardized tests are applied. The author's data indicate the test to be rather reliable.

Whether or not this or any other test in chemistry measures all the functions of chemistry is a matter of opinion. It may be said, however, that it does measure rather well the two objectives which most readily yield to measurement: namely, the acquisition of information and ability to solve representative problems.

RICH CHEMISTRY TESTS GAMMA AND EPSILON

These tests, having 25 multiple-choice questions each, were devised to measure achievement in general chemistry. The questions are intended to measure information, ability to solve numerical problems, ability to think with the materials of chemistry, and "habits and knowledge acquired from work in the laboratory." The working time is only twenty-five minutes. In this short period and with the relatively few questions, it is doubtful whether all these purposes can be achieved. The accuracy of the Rich tests is not so high as that of the Powers test; it is, therefore, less reliable.

Physics

IOWA PHYSICS TEST

Description of the test.—H. L. Camp devised three tests to measure the following branches of physics: (1) mechanics, forms 1 and 2; (2) heat, forms 1 and 2; (3)

electricity and magnetism, forms 1 and 2. Each form is given in a separate folder, some containing 11 questions and some 12. The questions are largely factual, and the answers thereto are to be written briefly on the examination booklet. No other materials are necessary.

The material of the tests is based upon 102 principles which were found by Starch to be common to five widely used textbooks of high school physics.

Scoring the test.—The time allowance is forty-five minutes for the tests of heat and mechanics and forty minutes for electricity and magnetism. In scoring, each problem is given a weighted value, and the final score is the sum of these values. As already stated, there is no advantage in employing a method of weighted scores, for statistical studies carried out to examine the merits of weighting have demonstrated that simple addition of the number of correct items yields almost identical results and is at times even more reliable.

Function of the tests.—These tests are intended to measure knowledge of the fundamental principles of physics and ability to employ this knowledge in the problems of "ordinary life." They provide an objective measure of a portion of the fields of physics, although it is likely that the test could be made much more searching if the exercises were in the multiple-choice form, which would permit the inclusion of more numerous and diversified questions. However, for the materials which the Iowa tests do cover, the reliability is quite high.⁴

⁴ We are indebted to Professor G. M. Ruch who furnished valuable statistical data bearing on a number of physics tests.

**Two QUESTIONS FROM THE IOWA PHYSICS TEST
ELECTRICITY AND MAGNETISM***Series C. Form 1. By Dr. Harold L. Camp*Value
(9.4)

6. What property of a volt-meter prevents it from short-circuiting the two lines when connected across?

ANSWER

(10.2)

7. What property of an electric current is utilized in comparing currents by means of a galvanometer?

ANSWER**HUGHES PHYSICS SCALE**

Description of the scales.—The Hughes Scales are four in number, two for the measurement of information, and two for the measurement of thought. Each scale consists of thirty exercises arranged in three groups, each of which is more difficult than the preceding group. The materials of these scales were selected on the basis of reports from teachers of physics, the reports being supplemented by analyses of textbooks.

Function of the scales.—The Hughes Physics Scales were built on a rather rigorous statistical basis. The scales are so arranged as to yield a score which indicates the difficulty of the item which the pupil is capable of handling with a correctness of 50 per cent. From the fact that the scales are separated into divisions treating information and thought problems separately, it is clear that they are designed to measure ability to employ a knowledge of physics, as well as to measure the extent of one's fund of information.

COLUMBIA RESEARCH BUREAU PHYSICS TEST

Description of the test.—This test is designed to do in physics what the Columbia Research Bureau Plane Geometry Test is designed to do in that subject. The physics test covers the materials common to the widely used high school and elementary college textbooks. The distribution of items is in approximately the following proportions:

Mechanics 16%	Light 16%
Heat 16%	Electricity 32%
Sound 8%	Miscellaneous 12%

There are two forms, A and B, each of which includes 144 questions of the "true-false" variety, arranged in order of difficulty. Among these 144 questions are to be found exercises of both information and reasoning, "with a preponderance of weight on reasoning and problem solving."

Scoring the test.—The working time for each form is 75 minutes, to be completed in one sitting. A special

key has been provided to expedite the marking of the papers, making the scoring rapid and objective. The final score is the number right minus the number wrong. This method, it will be recalled, is frequently employed with the "true-false" type of question in order to compensate for guessing.

Percentile norms and suggested scores for assigning letter marks are provided.

Function of the test.—The authors suggest that the test may be used as a means of improving instruction, as a basis for the assignment of marks, as an aid in student counseling, and as a basis for the assignment of college admission credits.

The data presented by the authors, based on about 900 students, show the test to be of high reliability. This is no doubt due in part to the large number of items in the test. A recent investigation conducted with about 300 students shows a reliability index which is moderately high.⁵

Comparison of tests in physics.—A recent statistical study⁵ designed to indicate the degree of correspondence between the several tests in physics showed that their inter-relationships were far enough removed from unity⁶ to warrant the view that they are not measuring precisely the same functions. We should, of course, not expect the correlations to be perfect, for even the correlation between two *forms* of the *same* test or two sets of scores for the *same group* of students on the same test will not be perfect, due to inherent difficulties in the test itself and to chance factors which produce varia-

⁵ Data from Professor Ruch.

⁶ For an explanation of correlational unity see Chapter XVIII.

tions in an individual's score from day to day. Yet the somewhat low⁷ intercorrelations between the different tests can not be accounted for entirely on the basis of chance variations. They are to be explained by the variations in content and emphasis from test to test, and because the pupils tested in the experiment were perhaps better prepared for some of the tests than for others, by virtue of the fact that their course of instruction was in closer agreement with those tests than with others.

One implication of these results for the teacher is that that test should be selected which, upon examination, seems best to meet the character of the course of instruction received by the students to be tested.

Other tests in secondary school science.—Space does not permit the presentation in detail of other highly regarded tests. They will, therefore, be briefly indicated here.

The Gerry Test of High School Chemistry is based on items selected from the examinations in chemistry of the College Entrance Examination Board, from 1911 to 1920. These items were checked with four textbooks.

The Glenn-Welton Instructional Tests in Chemistry contain thirty-six tests (in one booklet) covering every topic in high school chemistry. The number of items in each test varies from 20 to 80. They represent an interesting effort to analyze and measure in great detail the secondary school course of study in chemistry. They may be used every week or ten days, on the completion of a unit.

⁷ From +.57 to +.68; when corrected for attenuation, due to errors of measurement, the coefficients of correlation range from +.70 to +.86.

The Glenn-Obourn Instructional Tests in Physics consist of a series of twenty-five tests, each one covering a unit of a first course in physics in high school or college. They are to be used in the same manner as the Glenn-Welton tests in chemistry.

The Ruch-Cossman Biology Test is intended to measure general biology after one semester or one full year of instruction in the subject. The test has five parts: general biological information, incomplete statements, identification of structures from drawings, laws of Mendelian inheritance, and completion exercises.

The Michigan Botany Test is regarded by many as a good test in that subject. It is largely informational, although one group of questions (IV) contains many thought problems. Although the test requires but twenty-five minutes, the reliability appears to be rather high, judging from available data.

Conclusions.—Although the tests of secondary school science have their limitations and faults, they are, nevertheless, among the best measuring instruments of the high school group. There is in each of the sciences a large body of well organized materials which lend themselves readily to testing. The problem, of course, is the proper selections of essentials and their correct evaluation. Furthermore, the "broader aims," the habits, skills, and the "scientific attitude" are not at present being measured by the available tests. To be sure, these categories are difficult of mensuration; but this difficulty is in part due to the fact that those responsible for the teaching of the sciences are themselves not in agreement; nor have the categories been adequately defined.

One may seriously question whether it is the function of a test in the sciences to measure "broader aims," etc. Is not all education conceived with certain contingent and less immediate values in view? And if these contingent and less immediate values are to be achieved *through* the study of certain materials, is it not the first purpose of a test to measure the degree to which the necessary materials have been mastered? To measure the broader categories would, in fact, be to measure in part the development of personality through education. That is not the primary function of an achievement test.

For the most part, the tests described in this chapter measure rather satisfactorily that which they purport to measure.

MATERIALS NEEDED

Columbia Research Bureau Physics Test. (World Book Company, Yonkers-on-Hudson, N. Y.) Package of 25 tests with key and directions, \$1.30; specimen set, 25 cents.

Downing Range of Information Test. E. R. Downing, University of Chicago. Directions and scoring sheet, 10 cents; tests, 40 cents per hundred.

Dvorak General Science Tests. (The Public School Publishing Company, Bloomington, Ill.) Package of 25 tests, 50 cents; specimen set, 20 cents.

Gerry Test of High School Chemistry. (Ginn and Company, Boston.) 36 cents per 30.

Hughes Physics Scales (The Public School Publishing Company, Bloomington, Ill.) Package of 25 tests, 50 cents; specimen set, 15 cents.

Iowa Physics Tests. (The Public School Publishing Company, Bloomington, Ill.) Package of 25 tests, 50 cents; specimen set, 15 cents.

Michigan Botany Test. (The Public School Publishing Com-

pany, Bloomington, Ill.) Package of 25 tests, \$1.00, specimen set, 15 cents.

Powers General Chemistry Test. (World Book Company, Yonkers-on-Hudson, N. Y.) \$1.10 per package of 25; specimen set, 20 cents.

Rich Chemistry Test for High Schools. (The Public School Publishing Company, Bloomington, Ill.) \$1.00 per package of 25; specimen set, 20 cents.

Ruch-Cossman Biology Test. (World Book Company, Yonkers-on-Hudson, N. Y.) \$1.30 per package of 25; specimen set 10 cents.

Ruch-Popenoe General Science Test. (World Book Company, Yonkers-on-Hudson, N. Y.) \$1.30 per package of 25; specimen set 20 cents.

Van Wagenen Reading Scale in General Science. (The Public School Publishing Company, Bloomington, Ill.) \$3.00 per hundred; specimen set, 20 cents.

SUPPLEMENTARY LIST OF TESTS

Blaisdell Instructional Test in Biology.⁸

Columbia Research Bureau Chemistry Test.⁸

Cooprider Information Exercises in Biology.⁹

Denver Curriculum Semester Tests in General Science.¹⁰

Harvard Elementary Physics Test.¹⁴

Iowa Placement Examinations, Revised, Chemistry.¹¹

Iowa Placement Examination, Revised, Physics.¹¹

Michigan Instructional Tests in Physics.⁹

Peters-Watkins Objective Tests for High School Physics.¹²

Powers General Science Test.¹³

⁸ World Book Company, Yonkers-on-Hudson, N. Y.

⁹ The Public School Publishing Company, Bloomington, Ill.

¹⁰ Denver Public Schools, 414 Fourteenth St., Denver.

¹¹ Bureau of Educational Research and Service, University of Iowa, Iowa City.

¹² C. J. Peters, University High School, Columbia, Mo.

¹³ Bureau of Publications, Teachers College, Columbia University, New York City.

¹⁴ Ginn and Company, Boston.

Rauth-Foran Chemistry Tests.¹⁵

Starch Physics Test.¹⁶

Thurstone Vocational Guidance Tests—Physics.⁸

SELECTED REFERENCES

Camp, H. L., "An Evaluation of Standard Tests and Suggested Uses in Improving Physics Teaching" (*School Science and Mathematics*, Vol. 23, May, 1923, pp. 441-446).

Camp, H. L., "Scales for Measuring Results of Physics Teaching" (*Journal of Educational Research*, Vol. 5, May, 1922, pp. 400-405).

Cornog, J., and Stoddard, G. D., "Predicting Performance in Chemistry" (*Journal of Chemical Education*, Vol. 2, August, 1925, pp. 702-708).

Curtis, F. D., *Some Values Derived from Extensive Reading of General Science* (Contributions to Education, No. 163, Teachers College, Columbia University, 1924).

Downing, E. R., "The Revised Norms for the Range of Information Test in Science" (*School Science and Mathematics*, Vol. 26, February, 1926, pp. 142-146).

Dvorak, A., "A Study of Subject Matter and Achievement in General Science" (*General Science Quarterly*, November, 1925, January, March, May, 1926).

Glenn, E. R., "Bibliography of Science Teaching in Secondary Schools" (*Bureau of Education, Bulletin No. 13, 1925*; Department of the Interior, Washington, D. C.).

Powers, S. R., "Achievement in High School Chemistry—An Examination of Subject-matter" (*School Science and Mathematics*, Vol. 25, pp. 53-62; 1925).

Powers, S. R., "Tests of Achievement in Chemistry" (*Journal of Chemical Education*, Vol. 1, pp. 139-144, 1924).

Ruch, G. M. and Cossman, L. H., "Standardized Content in High School Biology" (*Journal of Educational Psychology*, Vol. 15, No. 5; pp. 285-296, 1924).

¹⁵ Catholic Education Press, Bookland Station, Washington, D. C.

¹⁶ C. A. Gregory Co., 345 Calhoun St., Cincinnati.

Windes, E. E., and Greenleaf, W. J., "Bibliography of Secondary Education Research, 1920-1925" (*Bureau of Education, Bulletin No. 2, 1926*; Department of the Interior, Washington, D. C.).

CHAPTER XV

FOREIGN LANGUAGES

Objectives in the study of foreign languages.—In perhaps no other subject of the secondary school has there been such a thorough and extensive systematic investigation of materials, methods, and results as in modern foreign languages. This is attested to by the seventeen volumes of the American and Canadian Committees on Modern Languages,¹ including volumes of word-books, studies of method and teacher training, use of achievement and prognosis tests, and a study of objectives. The American Council Tests, described later, are an outgrowth of what is known as the Modern Foreign Language Study conducted by the above named committees. In one of these volumes, V. A. C. Henmon, the author, states that “the four immediate objectives of instruction in the foreign languages are development of the ability to read, to write, and to speak the language and to understand it when spoken.”² To measure these abilities, however, it is necessary to break them into specific, measurable aspects. The following battery is therefore suggested: (1) a vocabulary test; (2) a silent reading test; (3) a translation-into-English test; (4) a translation test; (5) a written composition

¹ Published by the Macmillan Company.

² *Achievement Tests in Modern Foreign Languages*, by V. A. C. Henmon. The Macmillan Company, 1929, p. 3.

scale; (6) a grammar test; (7) an aural comprehension test; (8) a pronunciation test; and (9) an oral composition test. Of course, complete batteries for French, German and Spanish have not been developed; nor is it unlikely that the development of a test of pronunciation and oral composition will be found to be impossible. The other seven are capable of development, but thus far only part of the group has been constructed for each language.

For the study of Latin, very detailed objectives have been set down in Part I of the Classical Investigation.³ Among the numerous objectives there are listed such simple goals as the ability to read new Latin after the study of Latin has ceased, the ability to understand Latin words, phrases, abbreviations, and quotations occurring in English; and such complex, distant, and tenuous aims as the development of an appreciation of the literary qualities of the Latin authors read and the development of right attitudes toward social situations. Between these extremes will be found aims touching on proficiency in English, proficiency in other foreign languages, correct habits of thinking, knowledge of mythology, knowledge of Greek and Roman history, and the improvement of certain psychological functions. It is clear that the objectives to be achieved in the study of Latin are far more ambitious and lend themselves much less to measurement than the relatively modest aims of modern foreign languages as stated in Henmon's volume of the Modern Foreign Language Study. The available tests of Latin have, of course, had to con-

³ American Classical League, *The Classical Investigation*, Part I, 1924, pp. 38-79.

fine themselves to the more concrete aspects of the subject.

Modern Languages

AMERICAN COUNCIL ALPHA TESTS IN FRENCH, GERMAN, SPANISH, AND ITALIAN

Description of the tests.—In French, German, and Spanish tests have been constructed for the measurement of vocabulary and grammar (Part I), and silent reading and composition (Part II). There are two forms of each. In Italian, there is thus far an experimental edition of tests in vocabulary and grammar only, and in one form.

“The French vocabulary test consists of 75 words chosen from the successive groups of 50 words in a list of 3,905 words arranged in order of frequency of occurrence on a basis of a word count of 400,000 running words.”

“The German vocabulary test consists of 100 words selected systematically from Kaeding’s *Häufigkeitswörterbuch der deutschen Sprache* (1898), a word count based on 10,910,777 running words.” Before the vocabulary test was constructed, it was necessary to reduce this enormous list to a dictionary basis in order to determine the word frequency.

“The Spanish vocabulary test consists of 100 words chosen from a list of 6,702 words arranged in order of their importance on the basis of a word count of 1,200,000 running words.”

“The Italian vocabulary test consists of 100 words based on words common to beginners’ texts.”⁴

⁴ Henmon, *op. cit.*, pp. 8 ff.

In order to insure objectivity, as far as possible, the "recognition" technique was employed, using the multiple-choice form, as in the following:

1. *mais* hand more but month day
2. *prendre* approach take run hang paint

In *grammar*, the tests are functional in character, having, in each case, fifty items chosen from those which are common to widely used texts. The items are scaled in difficulty on the basis of the per cent of correct responses for each in preliminary investigations. The technique varies, though in French and Spanish the completion form predominates; whereas in German the recognition (or selection) type is used exclusively, and in Italian the completion form is employed throughout. For example, in the first case, completion, the pupil must write in missing definite articles, personal pronouns, etc.; while in the second, recognition, the pupil must designate which one of several sentences correctly translates a given English sentence.

In order to measure comprehension of silent reading, tests in French, German, and Spanish were constructed on the same plan as the Thorndike-McCall reading tests in English.⁵ The paragraphs, of increasing difficulty, call for answers in English, since the authors believe that comprehension of the passage read is better indicated than would be the case if answers were to be in the foreign language. The paragraphs were selected from a large number, and their order of difficulty was

⁵ See Chapter VII of this text.

determined by the percentage of correct responses for each.

On the last page of the reading comprehension test is a picture about which the pupils are to write a composition in the foreign language. The French scale was prepared by Professor M. R. Trabue, the German scale by Miss Elizabeth Rossberg of Milwaukee-Downer College, and the Spanish scale by Professor Henmon. In each case, sample compositions with assigned values are given, and pupil compositions are to be compared with these, much the same as in the case of English compositions.⁶

Function of the tests.—These tests offer a set of instruments with which the teacher of a modern language is enabled to measure with marked accuracy the four aspects of modern language ability, as indicated above. Extensive statistical studies have demonstrated that the batteries of tests meet the accepted standards of reliability.⁷

It must be remembered, however, that a test might very well yield high indices of reliability (consistency or stability of results) and yet fail to measure the ability which it purports to measure. The latter is known as the validity of a test. That is, how far does the test agree with the accepted criteria of the ability? Obviously, the criteria of validity must be determined by those who are engaged in the teaching of a subject and in the formulation of its aims and purposes. In the case of the American Council Tests herein described, it seems that the

⁶ See Chapter VIII, of this text.

⁷ Henmon, *op. cit.*, Chapter V.

items included are valid samplings, as indicated by the methods of selection and by the judgment of competent teachers who state that the tests reflect current practices.

AMERICAN COUNCIL BETA TESTS IN FRENCH, GERMAN, AND SPANISH

In addition to the Alpha tests, there are the Beta tests, which are on a somewhat lower level than the former, and to be preferred, perhaps, for the testing of first-year pupils. The test for each language has three parts: vocabulary, comprehension and grammar.

On the whole, the American Council Tests in modern foreign languages appear to be the most adequate of the measures in high school subjects.

COLUMBIA RESEARCH BUREAU TESTS IN FRENCH, GERMAN, AND SPANISH

Description of the tests.—Each of these tests is made up of three parts: a vocabulary test, a comprehension test, and a grammar test. There are two forms for each language, and all tests and forms are constructed upon the same principles.

In the *vocabulary* tests for each of the three languages there are 100 words, each of which is followed by four or five English words. The student indicates his knowledge of the word by underlining the correct English translation.

In each case the *comprehension* test consists of seventy-five statements in the foreign language. These statements, graded as to difficulty, make assertions

“which are obviously true or obviously false.” The student indicates his comprehension of the assertion by marking each one with a plus or a minus sign, as the case may be.

Part III, the *grammar* test, in all cases consists of 100 English sentences, each of which is followed by an incomplete translation. The student must complete the translations, the required completions being of such a nature as to examine his knowledge of grammatical forms.

The three parts of the tests in each language are based upon analyses of a number of representative text books. After the preliminary selection of materials, the items were used in three experimental editions. Out of these grew the final test.

Function of the tests.—These tests measure some of the specific objectives in modern language instruction. To be sure, the objectives measured are the mechanics of the language. They do not measure the cultural, oral, and aural aspects of language instruction. But, as already suggested in connection with other tests, the objective measurement of cultural aims in the case of any subject is perhaps impossible. Furthermore, we have also suggested that these cultural and broad objectives must rest upon a foundation of information, facts, or mechanics. It is altogether probable that the achievement of the “higher” aims of instruction in modern language will be facilitated by achievement in the language itself. The measurement of the latter, it seems, is therefore the first function of the standard, objective test.

Latin

ULLMAN-KIRBY LATIN COMPREHENSION TEST

Description of the test.—This test consists of ten paragraphs in Latin, each paragraph being followed by three or four questions in English which are to be answered in English. There are two especially constructed elementary paragraphs, four selections from Caesar, two from Cicero, and two poetical selections. The test, which comes in two forms, was used in the Classical Investigation.

Scoring the test.—The score is the number of questions correctly answered, the test and the scoring being similar to the Thorndike-McCall reading tests.⁸ Correct answers and variant forms are included in the scoring key, but all possible correct answers are not included, so that the scoring of answers becomes in part a matter of judgment. This, of course, lowers the objectivity of the measure. Norms are provided for eight semesters.

Function of the test.—The test is useful for the purpose of measuring reading ability in Latin and for the sectioning and comparison of groups on that basis. For individual diagnosis, however, it must be supplemented. The Classical Investigation found that the Ullman-Kirby Test as a measure of comprehension showed a high correlation with the test as a measure of translation.⁹

The reliability of the test is only fair; but this is possibly due in large part to the subjectivity which creeps into the scoring.

⁸ See Chapter VII of this text.

⁹ General Report, Part I, p. 194.

WHITE LATIN TEST

Description of the test.—The test consists of two parts. Part I contains a vocabulary of 100 words selected on the basis of frequency of occurrence in works read for college entrance. Part II consists of 20 Latin sentences arranged according to increasing difficulty of syntax. Both parts are in multiple-choice form. There are two forms, A and B.

Scoring the test.—The form of the test makes the scoring altogether objective. The score for Part I is the number of words correctly marked, while that for Part II is the number right multiplied by 5. Tentative norms (medians) are given for eight semesters.

Function of the test.—The White Latin Test will give a fair measure of vocabulary and translation for purposes of class-sectioning. The authors state that the test is designed for growth in knowledge of Latin through four years of study. This test, however, does not show sufficiently high reliability to warrant its use for individual measurement and diagnosis. It is useful principally for the study of groups.

Other Latin tests.—The Godsey Latin Composition Test, in two forms, was developed in connection with the Classical Investigation of the American Classical League. The test consists of three sections, each of which has eleven English sentences, followed by four Latin translations, only one of which is correct. Following each section is a group of sixteen rules, eleven of which cover the constructions of the preceding sentences. From four given rules (the number of the rules being stated after each sentence) the pupil is required

to designate the one which applies to the sentence, in order to justify his selection. The test is intended to be diagnostic.

The Pressey Test in Latin Syntax is made up of thirty-three English sentences with Latin translations in multiple-choice form. Only one translation of each sentence is correct, and the student is required to underscore the proper one.

The Tyler-Pressey Test in Latin Verb Forms consists of thirty-two Latin verb forms, each of which is translated into four English verb phrases. Only one of these is correct.

The Orleans-Solomon Latin Prognosis Test is distinguished from the others herein described inasmuch as it is intended to predict what success a pupil may be expected to have in the study of Latin. The test presents to the pupil a series of tasks involving simple learning in Latin, such as he would meet in the actual study of the language. The test has nine parts, as follows: (1) test of English vocabulary through derivations; (2) appreciation (recognition) of change in ending for gender and number of individual words; (3) recognition of change in ending for case in sentences; (4) use (by translation into Latin) of case endings in sentences; (5) recognition of change in ending for number (nouns and verbs) in sentences; (6) use (by translation into Latin) of number, case and verb endings in sentences; (7) distinction (by translation into Latin) between use of dative of indirect object and accusative with *ad* (place to which) in sentences; (8) test of volitional memorization of vocabulary; (9) test of vocabulary

recall (of words frequently used in the test as a whole).

The Henmon Latin Tests consist of fifty Latin words arranged in order of increasing difficulty. The student is to write after each word its meaning; and each word is given a definite weight or value, so that the score is the sum of the values of words defined correctly. The first, *bellum*, is given the lowest value, .4, and the last, *quisque*, is given the highest, 4.7. The steps between words vary from .1 to .3. In several cases, words were assigned equal values. This test represents seven years' study and experiment to select and to determine the relative weights of the words. They are taken from thirteen beginning Latin books and are all words in frequent use in the high school Latin course. There are two forms, for alternate use, to overcome practice effect and unfairness.

A second part of the Henmon tests—printed on the same sheet—consists of ten Latin sentences which are to be translated into English. These sentences were devised from the selected vocabulary.

In the field of Latin, there are, in addition to measures of the language itself, tests of historical background and classical allusions. Among these may be mentioned the following: the Davis-Hicks Test in Roman History, Late Republican Period; the Davis-Hicks Test in Historical Content and Background of Caesar's Gallic War; the Clark-Ullman Test of Classical References and Allusions.

Teachers of Latin at times find it desirable to determine the pupils' English vocabulary and knowledge of

forms. For this purpose the tests described in Chapter VIII will be found suitable.

Conclusions.—There can be little doubt that the researches carried out under the American and Canadian Committees on Modern Languages and by other investigators have resulted in a clearer formulation of aims in the study of modern languages and in the construction of tests which satisfy these aims to a marked degree. Of course, the tests are not perfect; nor do they measure the "higher" and less immediate aims; but they do measure those aspects which lend themselves more readily to measurement, and they do help in the establishment of uniform and objective standards of achievement. They also may function in making more nearly accurate analyses of the importance and difficulties of various language forms, with a view to improved teaching.

In Latin the investigations have not been so extensive. But there, too, much has been done in the way of more accurate measurement and evaluation. The teacher will do well, however, to use a "battery" of tests in Latin if it is desired to gain insight into a pupil's general achievement in the subject. This is necessary because the reliability of the Latin tests is not so high as one could wish, and because correlations between the various tests are sufficiently far removed from unity to indicate that they are not measuring the same functions throughout.¹⁰ In other words, they are similar only in part.

¹⁰ Brueckner, L. J., "The Status of Certain Basic Latin Skills," (*Journal of Educational Research*, Vol. 11, May, 1924, pp. 390-402).

In addition to the utilization of the tests, the teacher will find that the *Classical Investigation* (1924) of the Classical League contains many specific aids for the better teaching of Latin and for the better measurement of objectives.

MATERIALS NEEDED

American Council Alpha Tests in French, German and Spanish. (World Book Company, Yonkers-on-Hudson, N. Y.) Package of 25 tests, Part I, with directions and scoring key, \$1.25 (French and Spanish), \$1.30 (German); Part II, \$1.25, all. Specimen sets, French and Spanish, 35 cents; German, 40 cents.

American Council Beta Tests in French, German and Spanish. (World Book Company, Yonkers, N. Y.) Package of 25 tests, \$1.30; specimen sets, 25 cents.

Clark-Ullman Test on Classical References and Allusions. Bureau of Educational Research and Service, University of Iowa, Iowa City.

Columbia Research Bureau Tests in French, German and Spanish. (World Book Company, Yonkers, N. Y.) Package of 25 tests, each form, \$1.30; specimen set, 20 cents.

Davis-Hicks Test on the Historical Content and Background of Caesar's Gallic War. E. E. Hicks, Wilkinsburg, Pa.

Davis-Hicks Test in Roman History. E. E. Hicks, Wilkinsburg, Pa.

Godsey Latin Composition Test. (World Book Company, Yonkers, N. Y.) Forms A and B, \$1.00 per 25.

Henmon Latin Tests. (World Book Company, Yonkers, N. Y.) Tests, 1, 2, 3, 4 and X, each 50 cents per 25; specimen set, 10 cents.

Orleans-Solomon Prognosis Test. (World Book Company, Yonkers, N. Y.) Package of 25 tests, \$1.30; specimen set, 15 cents.

Pressey Test in Latin Syntax. (The Public School Publishing

Company, Bloomington, Ill.) Package of 25 tests, 50 cents; specimen set, 10 cents.

Tyler-Pressey Test in Latin Verb Forms. The Public School Publishing Company, Bloomington, Ill.) Package of 25 tests, 50 cents; specimen set, 10 cents.

Ullman-Kirby Latin Comprehension Test. Bureau of Educational Research and Service, University of Iowa, Iowa City. Forms 1 and 2, each \$1.75 per 100.

White Latin Test. (World Book Company, Yonkers, N. Y.) Forms A and B, each \$1.20 per 25.

SUPPLEMENTARY LIST OF TESTS

FRENCH

American Council French Grammar Test (Cheydleur).¹¹

Handschin Modern Language Tests—French.¹¹

Harvard French Vocabulary Test.¹⁴

Henmon French Tests¹¹

Iowa Placement Examination, Revised (French and Spanish).¹²

Sammartino-Krause Standard French Tests.¹³

Twigg French Vocabulary Test.¹⁴

Wilkins Prognosis Tests in Modern Languages (French and Spanish).¹¹

GERMAN

Van Wagenen and Hubman-Patterson German Reading Scale.¹³

¹¹ World Book Company, Yonkers-on-Hudson, N. Y.

¹² Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

¹³ The Public School Publishing Company, Bloomington, Ill.

¹⁴ Ginn and Company, Ashburton Place, Boston

LATIN

Deferrari and Foran Test in Latin Comprehension.¹⁷
Harvard Latin Test.¹⁴
Hutchinson Latin Grammar Test.¹³
Inglis Latin Tests.¹⁴
Lohr-Latshaw Latin Form Test.¹⁵
New York Latin Achievement Test.¹¹
Starch-Watters Latin Test¹⁶
Stevenson-Coxe Latin Derivative Test.¹³
Stevenson Latin Vocabulary Test.¹³

SPANISH

Contreras, Broom, and Kaulfers Test for Spanish Vocabulary.¹⁸
Contreras, Broom, and Kaulfers Silent Reading Test in Spanish.¹³
Stanford Spanish Tests.¹⁸
Wilkins Achievement Test in Spanish.¹⁹

SELECTED REFERENCES

Bagster-Collins, E. W., et al., *Studies in Modern Language Teaching (Publication of the American and Canadian Committees on Modern Languages, Volume 17)*, (The Macmillan Company, New York, 1930).

Bond, O. F., "Causes of Failure in Elementary Spanish and French Courses at the College Level" (*School Review*, Vol. 32, 1924, pp. 276-287).

Brueckner, L. J., "The Status of Certain Basic Latin Skills" (*Journal of Educational Research*, Vol. 9, May, 1924, pp. 390-402).

Churchman, P. H., "Courses for Beginners" (*The Modern*

¹⁵ Bureau of Educational Research, University of North Carolina, Chapel Hill, N. C.

¹⁶ D. Starch, 1374 Massachusetts Ave., Cambridge, Mass.

¹⁷ Catholic Educational Press, Washington, D. C.

¹⁸ Stanford University Press, Stanford, Cal.

¹⁹ Henry Holt and Company, New York.

Language Journal, Vol. 9, No. 4, January, 1925, pp. 207-225).

Clem, O. M., "Latin Prognosis: A Study of the Detailed Factors of Individual Pupils" (*Journal of Educational Psychology*, Vol. 16, March, 1925, pp. 160-169).

Coleman, A., "The First Year of Modern Foreign Language Study" (*The Modern Language Journal*, Vol. 10, April, 1926, pp. 389-399).

Coleman, A., *The Teaching of Modern Foreign Languages in the United States* (*Publication of the American and Canadian Committees on Modern Languages*, Volume 12), (The Macmillan Company, New York, 1929).

Fife, R. H., "Report on the Modern Foreign Language Study in the United States" (*Educational Record*, Vol. 6, July, 1925, pp. 203-211).

Henmon, V. A. C., *Achievement Tests in Modern Foreign Languages*. (*Publication of the American and Canadian Committees on Modern Languages*, Volume 5), (The Macmillan Company, New York, 1929).

Henmon, V. A. C., "A French Word Book Based on 400,000 Running Words," *Bulletin No. 3, Bureau of Educational Research*, 1924, University of Wisconsin.

Henmon, V. A. C., "Standardized Vocabulary and Sentence Tests in French" (*Journal of Educational Research*, Vol. 3, 1921, pp. 81-105).

Henmon, V. A. C., *Prognosis Tests in Modern Foreign Languages*. (*Publication of the American and Canadian Committees on Modern Languages*, Volume 14) (The Macmillan Company, New York, 1929).

Jordan, A. N., "Prognosis in Foreign Language in Secondary Schools" (*School Review*, Vol. 33, September, 1925, pp. 541-546).

Stoddard, G. D., "Iowa Placement Examinations," *University of Iowa Studies in Education*, Vol. 3, No. 2, August, 1925.

Ullman, B. L., and Kirby, T. J., "A Latin Comprehension Test" (*Journal of Educational Research*, Vol. 10, November, 1924, pp. 308-317).

Ward, C. F., *Minimum French Vocabulary Test Book* (The Macmillan Company, New York, 1926).

West, A. F. (Chairman). *Classical Investigation* (Princeton University Press, Princeton, N. J., 1924).

Wood, B. D., "New York Experiments with New-Type Modern Language Tests," (*Publications of the American and Canadian Committees on Modern Languages*, Volume 1) (The Macmillan Company, New York, 1929).

Young, C. E., "French and Spanish in the High School" (*University of Iowa Extension Bulletin*, No. 93, September, 1923).

CHAPTER XVI

GENERAL ACHIEVEMENT TESTS

The place of the general achievement test.—Throughout the discussion of the various tests in the preceding chapters, the point was made that the teacher and administrator, in selecting a test or tests in any subject, should consider their own needs, what aspects of a subject they desire to measure, and the aspects measured by the several tests which come under the same name. These factors necessitate a careful inspection of subject-matter tests before one may be selected. Not all tests in the same subject are equally valuable; nor have all authors had the same purpose in mind in devising their measures. Methods of scoring and the interpretation of scores may vary from test to test; and it may happen that considerable difficulty will be encountered in obtaining comparable scores for the several measures or a single index for the group of tests.

In order to facilitate survey testing, therefore, there are what are known as general achievement tests. These usually contain items in a variety of subjects; but the scores may be comparable, the tests are uniformly administered and scored, and they offer a single, composite basis for rating and comparison.

A caution, however, is necessary. It frequently happens that the principal advantages of a general achievement test for purposes of survey are administrative:

that is, the number of comparative studies of various tests is considerably reduced, orders need be placed with only one publisher instead of with several, different types of scoring and evaluation need not be mastered, etc. But these considerations alone are not sufficient to justify the use of a single general achievement test for survey purposes, if a group of independent tests—to cover the several subjects—seem to meet the situation more adequately, from the point of view of completeness, objectives, and reliability. In other words, administrative expediency should not supersede technical superiority of the tests themselves. If, however, the survey test is technically as satisfactory as the others, then the teacher and administrator enjoy the added advantages stated above. We once more, therefore, return to the point that it is necessary to select a measure or measures on the basis of local needs and objectives, on the one hand, and the nature of the test itself, on the other.

The tests about to be described were devised as a means of affording a single set of measures for the proper classifications of pupils, primarily; and also for the purpose of individual rating and educational guidance.

NEW STANFORD ACHIEVEMENT TEST

Description of the test.—The New Stanford Achievement Test is the 1929 revision and extension of the battery of tests which first appeared in 1923. There are two examinations, the primary and the advanced. At present two equivalent forms of each are available;

but the authors state that additional forms will be published as needed.

The Primary Examination, for grades 2 and 3, consists of five tests: reading (paragraph meaning), reading (word meaning), dictation (spelling), arithmetic reasoning, and arithmetic computation. The test of paragraph meaning is in the form of a competition test; that is, in each sentence one or more omitted words are to be supplied by the pupil, the missing words being dependent upon the paragraph. The test of word-meaning is of the multiple-choice form. The dictation (spelling) test consists in reading to the pupils a series of sentences of increasing difficulty. The test of arithmetic reasoning consists of twenty problems of increasing difficulty, while in arithmetic computation there are twenty-five exercises involving all four fundamental processes.

The Advanced Examination, for grades 4 to 9, consists of ten tests, as follows: reading (paragraph meaning), reading (word meaning), dictation (spelling), language usage, literature, history and civics, geography, physiology and hygiene, arithmetic reasoning, and arithmetic computation. The first three of these are of the same type as the corresponding tests in the Primary Examination. The fourth, language usage, consists of seventy-four sentences designed to measure grammatical usage and correct choice of words. In each sentence the pupil must select the correct one of two forms. The literature test has eighty items of information, in multiple-choice form. In history and civics there are also eighty items, principally of the informational type. The test in geography, number seven, and in physiology

and hygiene, number eight, are also of the informational, multiple-choice type. Each has eighty items. Number nine, arithmetic reasoning, consists of forty problems of increasing difficulty. Arithmetic computation, having sixty exercises, ranges "from simple primary combinations through successive degrees of complexity to the type of mathematics usually taught in the ninth grade."

The materials of the New Stanford Achievement Test are based upon independent selection by the authors from textbooks and other sources, and upon previous selections, such as the Thorndike Word Book, the Ayres and Buckingham spelling lists, and the University of Iowa investigation of tests in the social studies.

Method of using the test.—This achievement test is published either as single tests in each of the subjects or as a composite battery in a single booklet. The working time required for the Primary Examination is thirty minutes, plus the amount necessary for the dictation exercise. The Advanced Examination requires a working time of two hours, plus the time necessary for the dictation. Because of its length, the Advanced Examination should be administered in two or three sittings, preferably three. The directions for giving and scoring are simple, and the scoring is objective.

On the second page of each examination booklet is a "profile chart" upon which are plotted the pupil's scores in the individual subjects and his total score. The chart is so arranged that it is possible to read off norms for educational age, chronological age, and school grade.

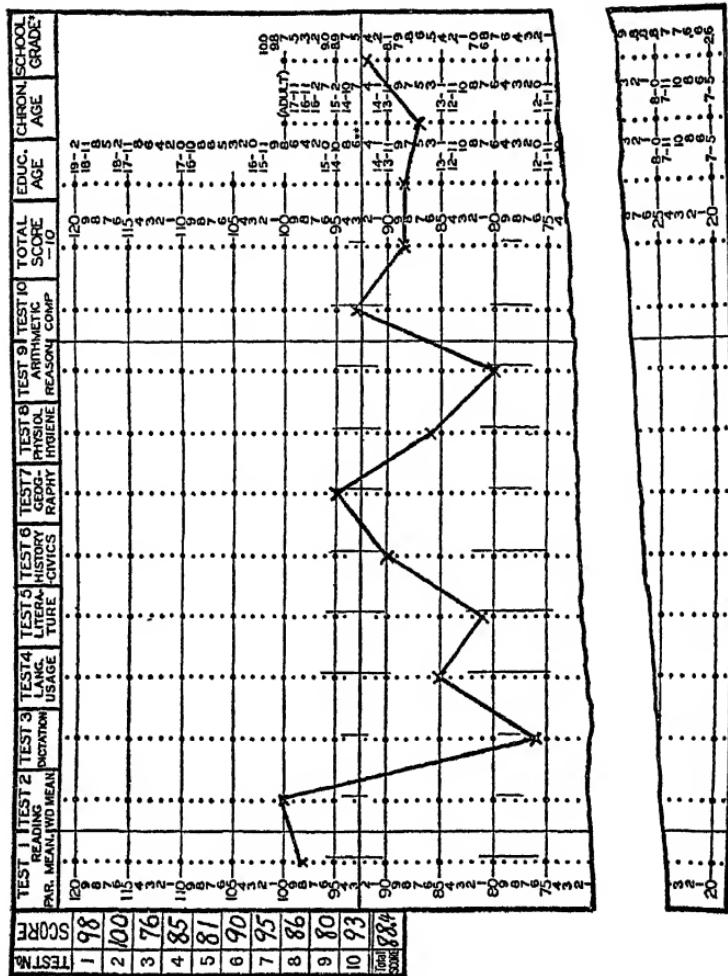


Fig. 2
PROFILE CHART FOR THE NEW STANFORD ACHIEVEMENT TEST

Function of the test.—This achievement test is intended for survey purposes in order to improve class groupings and for the study of individuals. One of the advantages of a battery of tests such as these is that they yield a composite score which may be trans-

DIRECTIONS: Draw a line under the word that makes the sentence true, as shown in the samples.

SAMPLES:

A rose is a
box flower home month river

A roof is found on a
book person rock house word

1 Ice is made of
baskets bread plants water wood

2 A castle is a
clock building path spirit wheel

3 Yesterday was a
day drive general heart tree

4 A maiden is a
bird boy girl king plant

5 A nest is a bird's
egg family food home tree

TEST 1. READING: PARAGRAPH MEANING

lated into the useful index, educational age (EA).¹ Furthermore, the use of the profile chart should prove of value in understanding an individual's variations in achievement from one subject to another and as compared with the norms of his age and grade.

Where the teacher or administrator is contemplating a rather wide survey of the several subjects, he will

¹ See Chapter III of this text for a discussion of the EA, the educational quotient (EQ), and the achievement quotient (AQ).

DIRECTIONS: Write JUST ONE WORD on each dotted line.

SAMPLE:

Dick and Tom were playing ball in the field. Dick was throwing the ball and.....
.....was trying to catch it.

1 Fanny has a little red hen. Every day the hen goes to her nest and lays an egg for Fanny to eat. Then she makes a funny noise to tell Fanny to come and get the.....

2-3 Anna had never seen a squirrel in her life, although she had always wanted to very much. One day when she was playing under a tree she heard a funny noise over her head. She looked up, and what do you think she saw? Up there in the.....was the very thing she had always wanted to see, a.....

TEST 2. READING: WORD MEANING

find the single, composite index of the achievement test valuable, provided the individual tests of the battery are themselves most suitable for the purpose at hand, and only then.

The authors' data show the *total scores* of the New Stanford Achievement Test to have high reliability for all grades. For some grades, however, the reliability of certain of the *individual* test scores is rather low.

DIRECTIONS: Draw a line under the word or phrase that makes the better sentence, as shown in the samples.

SAMPLES:

Apples is are good.

He told telled me.

1 He isn't any no better than you.

2 He doesn't don't know anything.

3 Please lay lie the book down.

4 He rang rung the bell.

5 I don't like them those apples.

TEST 4. LANGUAGE USAGE

SONES-HARRY HIGH SCHOOL ACHIEVEMENT TEST

Description of the test.—This test, having two equivalent forms, is devised for survey purposes in the high school. It covers four general fields which, the authors maintain, are usually required of all students in the secondary school: namely, language and literature, mathematics, natural science, and the social

studies. "The main attempt has been to sample as many parts of the four main fields of the secondary school core curricula as is consistent with secondary school testing conditions and reliable measurement." Under language and literature, for example, there are, among others, ten items on language usage, ten on word meaning, five on grammatical principles, ten on foreign phrases, five on character sketches, five on literary themes, ten on international authorship. In all, there are 140 items under sixteen types. In mathematics there are 80 items, including mathematical concepts, fundamentals of mathematics, geometric formulas, geometry theorems, mathematical formulas, and others. The third field, natural science, has 80 items also. These include processes in natural science, classification, principles, transformation of energy, instruments, and others. The fourth field, social studies, consists of 115 items, embracing, among others, civic information, American history, international affairs, geography, economic vocabulary and arguments.

The authors state that the range of difficulty is sufficient to measure reliability from grade 9B to "groups of practice teachers in the senior year of college."

Method of using the test.—The entire test is included in one booklet; but because the time limit for each of the four parts is forty minutes, it is clear that in the large majority of instances—if not all—it will be desirable to administer only one part at a sitting. The directions are simple, and the scoring is easy and entirely objective.

A profile chart is provided on the second page of each examination booklet. On this chart may be plotted

the individual student's percentile position in each of the four fields. Tentative percentile norms are provided for the separate tests for each half year and for grades 9B to 12A.

Function of the test.—The Sones-Harry High School Achievement Test is intended for use in grade and group placement and for the analysis of the course of study, the teacher, and the pupil. "It is based on the principle that the student as he advances through the secondary school should also be continuously adding to his fund of functional information." To derive a body of materials which would supposedly serve this purpose, the authors based their questions "as much as possible upon the agreement reached by various national committees and individual research workers in the various subject-matter fields." The questions were further revised upon the basis of criticisms of specialists, practice-teaching college seniors, and upon results obtained with 650 examinations. It is not at all clear how this method of selecting materials achieves the purpose of measuring a growing fund of functional information. Nor, it should be noted, is the test very thorough in any single branch of information measured. Any test is at best a sampling of a given subject or field of study. A single test, therefore, the scope of which is so broad as that of the Sones-Harry test, must of necessity restrict itself to reduced samplings if it is not to be inordinately long. These reduced samplings over a large area make a test useful only in a very general way.

The authors give no data with regard to the validity of the tests, that is, the extent to which they measure

what they purport to measure. However, the indices of reliability for all four fields are high, indicating that the test yields rather consistent results.²

Other general achievement tests.—Two others of the rather well known general achievement tests are the Otis Classification Test and the Illinois Examination. The former consists of a test of mental ability (Otis Self-Administering Test, Intermediate Examination) and an achievement test made up of 115 questions covering reading, spelling, grammar and diction, arithmetic reasoning and fundamental operations, geography, history and civics, physiology and hygiene, literature, vocabulary, music, art, and general information.

The Illinois Examination (I for grades 3, 4 and 5; II for grades 6, 7 and 8) is made up of three separate tests: Illinois General Intelligence Scale, Monroe's Silent Reading Tests (revised), and Monroe's General Survey Scale in Arithmetic. The three tests are simply arranged in convenient form to facilitate a testing program, and the "Teachers Handbook" indicates how the results may be combined for use in the school and class-room.

Conclusions.—One may very reasonably question the advisability of using general achievement tests, even in a survey, when a thorough measure and under-

² It is important to note that a test may have high reliability and at the same time have low validity. The index of reliability (self-correlation) simply indicates how consistent are the results obtained with the test. The index of validity indicates how well the test measures what it claims to measure; that is, how well it agrees with the criteria. High reliability alone is not sufficient to justify the selection of a subject matter test.

standing of pupil difficulties and achievements is desired. The general achievement test is by its very nature restricted in the number and variety of items it may contain; and its validity is therefore reduced. Consider, for example, the Otis Classification Test which touches sixteen different fields with only 115 items. The difficulties of constructing a general achievement test for high school use are even greater than in the case of the lower grades, for the fields of study have considerably expanded. Thus, in spite of a total working time of two hours and forty minutes, the Sones-Harry test is able to present only rather short and limited samplings in each part. The New Stanford Achievement Tests perhaps come closest to measuring general achievement with a reasonable degree of thoroughness. But even these tests do not show as high reliability for the several parts as may be expected. On the whole, it seems advisable to use the general achievement test only when a very general measure of accomplishment is desired. But if a survey of school achievement is to be searching and most reliable, if strength and weakness of pupil and school are to be analyzed, then it appears to be desirable to employ the best single tests in each subject to be measured.

MATERIALS NEEDED

Illinois Examination. (The Public School Publishing Company, Bloomington, Ill.) \$4.00 per hundred.

New Stanford Achievement Test. (World Book Company, Yonkers-on-Hudson, N. Y.) Primary Examination, package of 25 tests, \$1.10; specimen set (including guide), 50 cents. Advanced Examination, package of 25 tests, \$2.00; specimen set (including guide), 50 cents.

Otis Classification Test. (World Book Company, Yonkers-on-Hudson, N. Y.) Package of 25 tests, \$1.10; manual of directions, 25 cents; specimen set, 40 cents.

Sones-Harry High School Achievement Test. (World Book Company, Yonkers-on-Hudson, N. Y.)

SUPPLEMENTARY LIST OF TESTS

Indiana Composite Achievement Test.³

Primarily for grade 8.

Iowa High School Content Examination.⁴

For use near close of high school course, and for those entering college.

Iowa Placement Examinations, Revised.⁴

For students entering college; may be used in high school.

Lippincott-Chapman Classroom Products Survey Tests.⁵

Pressey Attainment Scales for Primary Grades.⁶

Grades 2 and 3.

SELECTED REFERENCES

Courtis, S. A., "The Influence of Certain Social Factors upon Scores in the Standard Achievement Tests" (*Journal of Educational Research*, Vol. 13, No. 5).

Kelley, T. L., Ruch, G. M., and Terman, L. M., "New Stanford Achievement Test" (Guide for Interpretation and Manuals). (World Book Company, Yonkers-on-Hudson, N. Y.)

Monroe, W. S., and Buckingham, B. R., "The Illinois Examination I and II" (Teacher's Handbook). (The Public School Publishing Company, Bloomington, Ill.)

Otis, A. S., "The Making of a Classification Test," *Contributions to Education*, Vol. 1, Chapter 14. (World Book Company, Yonkers-on-Hudson, N. Y.)

³ Bureau of Cooperative Research, Indiana University, Bloomington, Ind.

⁴ Bureau of Educational Research and Service, University of Iowa, Iowa City.

⁵ J. B. Lippincott Company, Washington Square, Philadelphia.

⁶ The Public School Publishing Company, Bloomington, Ill.

Pintner, R. and Marshall, H., "A Combined Mental-Educational Survey" (*Journal of Educational Psychology*, Vol. 12, January and February, 1921).

Sones, W. W. D. and Harry, D. P., "Sones-Harry High School Achievement Test" (Manual of Directions). (World Book Company, Yonkers-on-Hudson, N. Y.)

CHAPTER XVII

INTELLIGENCE TESTS¹

The nature of the intelligence test.—In Chapter III it was pointed out that the teacher can make good use of intelligence tests in the solution of some class-room and individual problems and for the purpose of diagnosis. It is true, however, that some psychologists are of the opinion that there is need of more expert skill and psychological training than most teachers have if test results are to be interpreted and used correctly. This limitation is particularly pertinent in the matter of individual testing and in diagnosing cases which belong in the psycho-educational clinic. But for the usual problems of the class-room and school where general guides and indices are necessary, the group test may be employed advantageously by the teacher after brief instruction in administering and scoring the tests and interpreting results.

Since the widespread use of intelligence tests there has been considerable discussion and controversy re-

¹ Among psychologists there is some objection to the use of the term "intelligence tests" for the type of measure to be described in this chapter. Some would prefer to call them tests of "mental ability" or simply "mental tests"; others prefer the term "school aptitude" tests. The objection to the term "intelligence tests" arises from uncertainty as to the nature and definition of intelligence. We have, however, retained the name "intelligence tests" because the term has come to signify a definite *type* of measure. Whether that measure actually is a true test of intelligence is another question, and one for the more advanced student to deal with.

garding the use of the tests and what it is they actually measure.² Much of the controversy has centered about the definition of intelligence and the adequacy of the tests for its measurement, as well as around the general question of "nature" and "nurture"; that is to say, how far the tests measure "innate" ability, and how far they reflect learning and training.³ Likewise there have appeared a variety of questions and problems, more or less theoretical, in connection with intelligence and its measurement. But, perhaps fortunately, the teacher need not be too much disturbed by these problems and controversies so far as the practical application of the tests is concerned; for whatever else the tests measure, there is little doubt that they do measure with a marked degree of reliability the *ability to do school work*. Whether that ability be the product of inheritance or environment, or both, is indeed a very important matter to teacher and theorist alike. But as we have already implied, the tests serve very well in identifying the kind of pupil material with which the teacher must work *at the time of measurement*;⁴ and that, it seems, is the

² See A Symposium, "Intelligence and its Measurement," (*Journal of Educational Psychology*, vol. 12, pp. 123-147, 195-216, 271-275).

³ For a variety of studies on this question consult the Twenty-Seventh Yearbook of the National Society for the Study of Education, volumes I and II. See also W. F. Dearborn, *Intelligence Tests*. (Houghton Mifflin Company, Boston, 1928.)

⁴ It should be added that the present view is that the IQ is constant within rather narrow limits; that is, an individual maintains his relative rank or degree of brightness rather consistently throughout his school life. This is true if there are no unusual circumstances; for it is also true that marked improvement in environmental conditions, including health and educational opportunity, will produce some changes in the IQ, just as changes for the worse in health and general environment may depress the IQ. Cf. "Intelligence Tests and the Nature-Nurture Controversy," by F. S. Freeman, *School and Society*, vol. 30, No. 782, pp. 830-830. Also W. F. Dearborn, *op. cit.*, Chapters IV and V.

significant fact for the teacher, supervisor, and administrator, in so far as the actual function of the test is concerned.

It should be remembered that tests of intelligence (or *mental ability*, as they are sometimes called) do not purport to measure an individual's personality, character, temperament, industry, or any other such qualities which contribute to success or failure, in school and out, in spite of what might be expected on the basis of intelligence ratings alone. At times teachers will discover from the results of an intelligence test that a pupil ranks very high, whereas his school work had led her to rate him as mediocre or poor. On the other hand, the tests may disclose the fact that the brightness of another pupil is more apparent than real; that he is, in fact, very ordinary or dull, but that he compensates for this by means of a ready enthusiasm, willingness, and a surface alertness. No formula can be stated for the handling of such cases. They require skillful consideration from a teacher experienced in such situations, or from a psychologist, either of whom recognizes the necessity of taking into account all of the individual's abilities and qualities.

Types.—Two types of tests are in use, the *individual* and the *group* tests. The former, as the name indicates, is given to one pupil at a time; and, as has already been stated, requires that the examiner shall have had special training. The test should be given, so far as possible, in a room where there are no distractions and where no one is present except the pupil and the examiner, except under unusual circumstances. Inasmuch as we maintain that the individual test should not be

administered by anyone who has not been specifically trained for the purpose, we shall not concern ourselves with the problems and technique of giving this type of test. We wish to make it clear, however, that where a pupil is a "problem case" it is advisable, wherever possible, to have him examined individually; for in such instances the score is more reliable than that of the group test, and the psychologist's report adds information which can not be ascertained when the group test is employed.

The *individual test* of to-day really owes its origin to the French psychologists, Binet and Simon, particularly Binet. During the last decade of the 19th century, Binet, like a number of other psychologists, had experimented with various sensori-motor and simple association tests as a means of measuring mentality. About 1900, however, he realized that if the tests were to be of value they must measure the complex mental processes rather than the simple sensory processes. With that principle in mind he conducted his experimental work and in 1905 produced the first scale. This scale, after trial and analysis, was revised and produced as the scale of 1908. This in turn was further subjected to experimentation and revision and appeared as the 1911 scale. Unfortunately, Binet died that same year, and his leadership in the work was lost. To him we owe the concept of the mental age (MA) and the concept of a series of questions and problems graded in difficulty.

The Binet-Simon scale has been revised and adapted for use in this country by Goddard, Herring, Kuhlmann, Terman and Yerkes. Of these, the revision per-

haps most widely used is the one known as the Stanford Revision of The Binet-Simon Tests, made under Terman. The Stanford Revision consists of ninety items (74 main questions plus 16 alternatives), arranged in groups of increasing difficulty from the age level of three years up to that of the "superior adult." Each age group contains six items, with the exception of year XII, which has eight. The tests include such items as weight discrimination, visual and auditory memory, counting, detecting absurdities, vocabulary, judgment, disarranged sentences.⁵

This test is regarded by many as a most reliable instrument for measuring learning ability, particularly of pupils of elementary and intermediate school age. For the measurement of mental ability below the age of 5 or 6 and above 15 or 16 there is doubt of its adequacy, as there is of the adequacy of most tests. The ages below 5 and above 15 present difficult problems which tests up to now have not met so well as they have those of the intermediate years, though it should not be thought that the tests for these intermediate years are in their final and completely satisfactory form.

The *group tests* may, of course, be given to an individual or to the entire class at one time, though they are intended primarily for group use. The size of the group which may be tested will depend on a number of considerations. One general specification may be made: The group should be such that conditions will be con-

⁵ For a discussion of the standardization and use of the Stanford revision, and for a description of the tests and their scoring, see Terman, L. M., *The Measurement of Intelligence* and *The Intelligence of School Children*.

ducive to order and complete control on the part of the examiner. The saving in time effected by the use of a group test is very great, for a class of 30 or 40, or more, can be examined in from 30 to 60 minutes, depending upon the test used, whereas an individual test requires from a half hour to an hour and a half for each pupil, depending upon the age, facility, and brightness of the subject.

Some representative group tests.—The group tests of intelligence have incorporated the principle of the Binet-Simon, with modifications, additions or eliminations here and there as experience has brought new facts to bear on the problems of mental measurement. The first of these to have widespread use was that of A. S. Otis who in 1917 prepared a group of tests to be used as a rough measure of intelligence. His work was made the basis of the form and, to some extent, the content of the tests constructed for use in the United States Army during the World War. After the war, Otis revised his earlier tests and published them in two parts, the Primary Examination for grades 1-4, and the Advanced Examination for grades 5-12. The Primary, requiring 25 minutes to give, is made up of eight tests, of which six are non-verbal. The Advanced, requiring 60 minutes, consists of ten parts, as follows: (1) Following Directions; (2) Opposites; (3) Disarranged Sentences; (4) Proverbs; (5) Arithmetic; (6) Geometric Figures; (7) Analogies; (8) Similarities; (9) Narrative Completion; (10) Memory. Both the Advanced and the Primary have two equivalent forms, so that on re-testing, familiarity with one form may be overcome by using the other, though the danger of

recall is not great if there has been no specific coaching.

The use of group tests in the army during the World War gave rise to a widespread utilization of similar tests in education and industry, with a consequent increase in their number as well as in their application to experimental problems in education and psychology. At the same time, the tests themselves have been subjected to experimental scrutiny so that one may now make selections from a list of tests which serve a very useful function. We shall describe only a few of those widely used in order to indicate the nature of the instruments.

The National Intelligence Tests were prepared by Haggerty, Terman, Thorndike, Whipple, and Yerkes. These consist of two scales of five tests each, Scale A involving Arithmetic Problems, Sentence Completion, Logical Selection, Synonym-Antonym, and Symbol Digit tests; and Scale B, Computation, Information, Vocabulary, Analogies, and Comparison. Five alternative forms of each scale were prepared. The test is for use in grades 3 to 8 inclusive and takes from 30 to 35 minutes to give. Though the results of one form may be used independently, the committee recommends that both forms be used and the results compared, so that in the event of a marked discrepancy an individual examination may be given.

The Dearborn Group Tests of Intelligence consist of two series, one for grades 1 to 3 and the other for grades 4 to 12. Series I, for grades 1 to 3, is made up chiefly of pictorial and geometric series, examinations A and B, both of which should be given if the most

reliable results are desired, although examination A may be used alone as an abbreviated form. In testing with Series I, A is given first, followed after an interval of a class period by B. The author of the test states, however, that for testing in the first grade it is often desirable to divide the tests into four parts, to which they readily lend themselves. Series I has no time limits, but should require no more than two periods. Series II, for grades 4 to 12, contains seven tests: (1) Picture Sequences; (2) Word Sequences; (3) Form Completion; (4) Opposite Completion; (5) Faulty Pictures; (6) Disarranged Proverbs; (7) Number Problems. This series also is in two parts, examination C and D, requiring from 30 to 35 minutes each. The Dearborn tests are distinguished from many others by the absence of emphasis on language ability; in fact, Series I is entirely non-verbal, while II places comparatively little emphasis on the verbal.

The Pintner-Cunningham Primary Mental Test is a well ranked test for the kindergarten, first, and second grades. There are seven parts to the test, all of which are composed of pictures. First, the pupil marks certain parts of the pictures as directed by the oral instructions of the examiner; second, he picks out the prettiest picture from two sets of similar pictures; the third part is an Associated Objects test; the fourth is a Discrimination-of-size test; the fifth is a Picture Parts test; the sixth is a Picture Completion test, and the seventh is a Dot Drawing test. The test requires from 30 to 50 minutes. The authors found it yielded a high coefficient of reliability; that is, the test gives

consistent results. They also found it closely related to the teacher's judgment of a pupil, but more nearly valid in rating intelligence.

The Haggerty Intelligence Examination, Delta I, is used in grades 1 to 3, and Delta 2, in grades 4 to 9. Delta 1 consists of six tests, of which five are non-verbal. It requires 30 minutes for giving. Delta 2 requires 35 minutes and is made up of the following six tests: Sentence Reading, Arithmetic, Picture Completion, Synonym-Antonym, Common-Sense, General Information.

The Mentimeters by Trabue and Stockbridge are intended for all levels of intelligence, from the first grade to the university. The time required varies, of course, with the group to which the test is being administered. In the Mentimeters are found Picture Absurdities, Maze Threading, Geometric Figures, Opposites, Reading Directions, Completion, Range of Information, Arithmetic. The Mentimeters are intended to measure academic ability, mechanical skills, and artistic judgment.

The Illinois Examination, by Monroe and Buckingham, is devised to measure the relationship between capacity and performance. The examination consists of two parts, one being a typical intelligence test, and the other a test in arithmetic and reading. In scoring, the two parts are kept distinct, so that for each pupil there will be a mental age and an achievement age. It is thus possible to derive an achievement quotient by means of the test, as well as the two ages. The Illinois Examination has two forms, I and II, the former being for grades 3, 4, 5, and the latter for 6, 7, 8. The time

required is about 60 minutes. The intelligence tests comprise: Analogies, Arithmetic Problems, Sentence Vocabulary, Substitution, Verbal Ingenuity, Arithmetical Ingenuity, Synonym-Antonym.

Other tests which are of the same type as the Illinois are the Mental-Educational Survey Test by Pintner, the Otis Classification Test, and the New Jersey Composite Test.

The Terman Group Test of Mental Ability is designed for grades 7 to 12. It consists of ten tests: Information, Best Answer, Word Meaning, Logical Selection, Arithmetic, Sentence Meaning, Analogies, Mixed, Classification, Number Series. It takes 35 minutes in application.

The Kuhlmann-Anderson Intelligence Tests are among the most recent in the field. They are equally well adapted to group testing and to individual examinations. The tests are divided into nine series, suitable for grades 1 to 12 inclusive. A separate series has been prepared for the first semester of grade 1 and another for the second semester of grade 1. There are also separate series for grades 2, 3, 4, 5 and 6; one also for grades 7 and 8, and another for grades 9 to 12. The nine series include a total of thirty-nine tests, containing, among other, Direction, Similarities, Disarranged Words, Geometric Figures, Opposites, Disarranged Sentences, Number Sequences, Arithmetical Ingenuity. The time required varies with the series being used.

There are also a number of group tests of intelligence intended for use with high school seniors and college freshmen. Although the class-room teacher will ordi-

narily not be concerned with these, they should, however, find place in a list of group tests. They include: the Thurstone Psychological Examination, the Brown University Psychological Examination, the Thorndike Intelligence Examination, the Morgan Mental Test, the Psychological Examination of the American Council on Education, and the Ohio State University Psychological Test.

Each of the group tests mentioned has its manual for giving and scoring. Ordinarily, too, the tests will be accompanied by stencils and other devices which facilitate the scoring of the papers. In addition, each psychological test generally has its descriptive booklet containing norms, statistics of validity and reliability, and an explanation of the purpose of the test. It is essential that the teacher or administrator become familiar with this descriptive booklet before the test is employed.

Performance and Aptitude Tests

It is necessary to distinguish between the several types of tests in use. We have thus far differentiated between the intelligence and the achievement tests. But we must indicate two further distinctions.

Performance tests.—The performance test is one in which the subject gives his response by *doing* something rather than in terms of language, as is commonly the case with the group test. Language is reduced to a minimum in the directions, or eliminated in some instances; and no language is necessary for the response. This type of test is especially valuable in testing those

with a language deficiency, such as the foreign subject, and also in examining the deaf. The performance scale is further valuable as a supplement to scales of the Binet and verbal group test types, in that they tend to compensate for the emphasis on language.

A common type of performance test is the *form board*. Depressions are cut in a board, and these depressions are to be filled by appropriate blocks. The geometric nature of the depressions varies, as do the difficulty and complexity of filling them. Among the well known tests of this sort are the Dearborn Form Board Tests and the Seguin Form Board.

A second type is the *maze* test. The subject is required to trace the correct route through a maze, which may vary in difficulty from the very simple to the highly complex. Of this type we may mention the Porteus Maze Test and the Dearborn Maze Tests.

Other performance tests include fitting together pictures, fitting together geometric figures, geometric figures in the nature of puzzles, putting together parts of a manikin or of a profile, picture completion, copying designs, and others. It is clear, of course, that these call for motor rather than for verbal responses. The most elaborate of the scales utilizing these motor tests are the Pintner-Paterson Performance Scale,⁶ the Army Performance Scale, the Arthur scale for advanced grades, and the Merrill-Palmer tests for preschool and first grade children. Performance scales have been used principally for examining the deaf and non-English speaking groups. For the school, their chief

⁶ For a description of this scale see *A Scale of Performance Tests*, by R. Pintner, and D. G. Paterson, (D. Appleton and Company, New York).

value lies in assisting in the detection of apparent disabilities which might be due to a language difficulty rather than to genuine retardation. Where time permits and where clinical facilities are adequate, it is desirable to administer a performance test in the diagnosis of an individual problem case.

Aptitude tests.—As its name suggests, the aptitude test is intended to measure a particular aptitude or group of aptitudes. In a sense the tests of intelligence which were described above are aptitude tests, inasmuch as they measure aptitude for school work; and the achievement tests described in the earlier chapters are likewise measures of aptitudes, in a sense, for they serve as indicators and forecasters of learning in those subjects. But ordinarily, when speaking of an aptitude test, we have in mind one of the non-academic type, such as a measure of mechanical ability, ability as an engineer, a clerk, a telegrapher, a typist, an aviator, a telephone operator, etc.; in other words, specific vocational ability. Examples of these are the Stenquist mechanical aptitude test, the Seashore test of musical ability, the Münsterberg test for motormen, the Army trade tests for mechanics of various types, the Thurstone vocational guidance tests, the Minnesota Mechanical Aptitude tests, and the Stanford tests of scientific aptitude.

The tests of various types mentioned in this chapter by no means offer a complete list of those in use. The selection has been such as to give a sufficient range and variety to enable the teacher or administrator to see the types of measures from among which selection may be made in dealing with school problems. Furthermore,

it has been the purpose of this chapter to indicate the nature of the materials incorporated in current tests.

MATERIALS NEEDED

Dearborn, W. F., Group Tests of Intelligence, Series I, grades 1-3. Series II, grades 4-12. (J. B. Lippincott Company, Philadelphia.) Specimen set 15 cents, either series 25 booklets \$1.50. Examiners' Guide 10 cents.

Haggerty, M. E., Intelligence Examination, Delta 1, grades 1-3. Delta 2 grades 8-9. (World Book Company, Yonkers-on-Hudson, N. Y.) Specimen set 55 cents, Delta 1 \$1.30 for 25 booklets, Delta 2 \$1.25 for 25 booklets, Manual of directions 25 cents.

Illinois General Intelligence Scale, Illinois Examination I for grades 3-5, Illinois Examination II for grades 6-8, forms 1 and 2 of each. (The Public School Publishing Company, Bloomington, Ill.) Sample set 20 cents, \$2.00 per hundred.

Kuhlmann-Anderson Intelligence Tests, for grades 1-12. The Educational Test Bureau, Minneapolis, Minn. Specimen Set (including forms for all grades) \$1.20. Package of 25 copies, \$1.25.

National Intelligence Tests. (World Book Company, Yonkers-on-Hudson, N. Y.) Scale A, form 1, Scale B, form 2, are all alternative forms. Any form complete \$1.30 for 25, Manual of Directions, 20 cents.

Otis, A. S., Group Intelligence Scale, Primary Examination for grades 1-4, Advanced Examination for grades 5-12, forms A and B of each. (World Book Company.) Specimen set 50 cents. Primary examination \$1.25 for 25, Advanced Examination \$1.30 for 25. Manual 30 cents.

Pintner and Cunningham, Primary Mental Test for Kindergarten to second grade. (World Book Company, Yonkers-on-Hudson, N. Y.) Specimen set 20 cents, \$1.45 for 25 including Manual.

Terman, L. M., (individual examination) for any age above 3 years. (Houghton Mifflin Company, Boston.) Condensed

Guide \$1.00, Test Material \$1.00, 25 Record Booklets \$2.00,
Abbreviated Filing Cards \$1.00 per hundred.

Terman, L. M., Group Test of Mental Ability for grades 7-12,
forms A and B. (World Book Company, Yonkers-on-
Hudson, N. Y.) Specimen set 15 cents. Either form com-
plete \$1.35 for 25.

Trabue and Stockbridge, Mentimeters for any grade. (Double-
day, Page & Co., Garden City, N. Y.) Specimen set 25 cents,
\$1.75 for 25 pupils complete.

SUPPLEMENTARY LIST OF TESTS

Cole-Vincent Group Test of Intelligence for School En-
trance.⁷ For kindergarten or first grade.

Detroit (Engel) First-Grade Intelligence Tests.⁸ For chil-
dren entering the first grade.

Detroit Mechanical Aptitude Examination.¹⁰

Miller Mental Ability Test.⁹ For grades 7-12 and college
freshmen.

Multi-Mental Scale for Elementary School.⁹ For grade 3 and
above.

Otis Self-administering Tests of Mental Ability.⁸ Intermediate
examination for grades 5-9. Higher examination for
grades 9-12.

Pressey Classification and Verifying Test¹⁰ Primary Test for
grades 1-3. Intermediate Test for grades 3-6. Senior Test
for grade 7 and above.

O'Rourke Mechanical Aptitude Test—Junior Grade.¹¹

SELECTED REFERENCES

Dearborn, W. F., *Intelligence Tests* (Houghton Mifflin Com-
pany, 1928).

⁷ Bureau of Educational Measurements and Standards, Teachers Col-
lege, Emporia, Kan.

⁸ World Book Company, Yonkers-on-Hudson, N. Y.

⁹ Bureau of Publications, Teachers College, Columbia University.

¹⁰ The Public School Publishing Company, Bloomington, Ill.

¹¹ Educational and Personnel Publishing Co., Washington, D. C.

Freeman, F. S., "Intelligence Tests and the Nature-Nature Controversy" (*School and Society*, Vol. 30, No. 782, pp. 830-836.)

Hull, C. L., *Aptitude Testing* (World Book Company, Yonkers, N. Y., 1928).

Kohs, S. C., *Intelligence Measurement* (The Macmillan Company, New York, 1927).

Levine, A. J. and Marks, L., *Testing Intelligence and Achievement* (The Macmillan Company, New York, 1926).

National Society for the Study of Education, *Twenty-first, 1922, and Twenty-seventh, 1928, Yearbooks* (The Public School Publishing Company, Bloomington, Ill.).

Pintner, R., *Intelligence Testing* (Henry Holt and Company, New York, 1923).

Pintner, R. and Paterson, D. G., *A Scale of Performance Tests* (D. Appleton and Company, New York, 1917).

Symposium, "Intelligence and its Measurement" (*Journal of Educational Psychology*, Vol. 12, pp. 123-147, 195-216, 271-275).

Terman, L. M., *The Measurement of Intelligence* (Houghton Mifflin Company, Boston, 1916).

Terman, L. M., *The Intelligence of School Children* (Houghton Mifflin Company, Boston, 1919).

Thomson, G. H., *Instinct, Intelligence and Character* (Longmans, Green & Co., New York, 1925).

Thorndike, E. L. et al. *The Measurement of Intelligence* (Bureau of Publications, Teachers College, Columbia University, 1926).

Young, Kimball, "The History of Mental Testing" (*Pedagogical Seminary*, vol. 31, pp. 1-48, March, 1924).

CHAPTER XVIII

STATISTICAL AND GRAPHIC METHODS

Widespread use of statistics.—In the course of our discussion up to this point we have had occasion frequently to employ statistical terms. In fact, the terms have been used throughout, particularly in the chapters dealing with the characteristics of tests, their validation and use, and in those chapters describing the tests themselves. In order that those who are unfamiliar with statistical indices and devices may have a better understanding thereof, it is the purpose of this chapter to present the simpler facts of statistics and of graphic methods in language untechnical as far as possible.

It is almost impossible to read current periodicals and publications in education and educational psychology without at least an understanding of statistical terms and devices. It is, therefore, particularly important that the teacher and school administrator have a grasp of the more common aspects of the matter, both for the purpose of conducting a program of standardized testing, and for the purpose of critical evaluation of their own and others' procedures and results. The use of statistical method in education and psychology has grown rapidly, especially since the widespread employment of tests of intelligence and of achievement in the schools. But, unfortunately, critical evaluation

has not kept pace in growth, with the result that studies have been carried on to show averages, deviations, correlations, etc., without a clear or markedly successful attempt to define these results in terms of the peculiar problems inherent in education or psychology. In other words, statistical method and *its* problems have often been mistaken for or confused with the problems of education and psychology. Our point is that statistics are extremely valuable as an *aid* and as an *additional source* of information; but we must go beyond the bare indices, if statistics are to serve their best purpose; for statistical findings must be viewed first with a clear understanding of the field in which they are being applied, whether it be education, psychology, sociology, or economics.

The final value of any program of measuring the results of teaching depends upon the information gained as a result of the testing. Such information can be had from an orderly presentation of the results obtained. Too often tests have been given with a feeling that there is virtue in the tests themselves; and, consequently, the results have been filed away without having been used. This has generally occurred because the teacher, supervisor, or administrator had not a clear conception of the problem being dealt with, or because they were unable to handle and interpret the data. We shall, therefore, consider the more common methods of analyzing results. The need for a well formulated program of testing has already been indicated.

Statistical and graphic methods.—There are two methods of presenting data: the statistical and the graphic. The science of statistics is the treatment of

numerical values, obtained by measurement, in such a manner as to present the results briefly and intelligently through precise indices. These same data which furnish the material for statistical treatment may also be presented by means of curves, or graphs, in such a way as to show the same facts and results in another form. Indeed, the two methods—statistical and graphic—should be used together wherever possible; for the graph offers a pictorial representation of numerical values which might be much less readily comprehended without the supplementing curve.

Statistical Indices

Complicated though some aspects of statistical method may be, the main purpose in most ordinary problems is to show one or more of three facts concerning groups of measures: namely, the central tendency, the deviation (variability or dispersion), and the index of correlation. These will be considered in the order named.

Measures of central tendency.—The central tendency is a measure of the “typical” case within the group; it is the value which most nearly represents the group. In standardized tests, for instance, it is the “norm.” It must be emphasized that the measure of central tendency applies primarily to the *group* being studied; it does not signify that any *individual* member of that group *necessarily* conforms with the central tendency. This fact, apparently ridiculously simple, is, nevertheless, frequently overlooked in reasoning and generalizing based on group data. For example, we may say that

the average IQ in a given grade is 105 and that the range is from 85 to 125; yet it is possible that not one pupil actually has an IQ of 105. It is clear from these simple facts that though the average represents a middle group fairly well and states the general trend of the group, there are individual pupils who fall below or exceed the average in varying degrees. In fact, we wish to make it clear that statistical indices are of value primarily as *characteristics* of groups, and as standards with which an individual may be compared. But measures of intelligence and of achievement are intended even more for the understanding of individuals than of groups. Thus, the psychology of individual differences must not be forgotten or swallowed up in a psychology of groups, valuable though the latter be.

There are three measures of central tendency in common use, (1) the arithmetic mean, (2) the median, and (3) the mode.

THE MEAN: The arithmetic mean (or average)¹ is the best known of the three and may be defined as the sum of the separate scores or measures in a group, divided by the number of measures. In a simple series it is necessary only to add up all the scores and divide by the number of measures. Thus, the pupils in a class score 65, 70, 75, 80, 85, 90, 95; the average will be 80 $\left(\frac{560}{7}\right)$.² But it nearly always happens when a large number of cases are being dealt with that the meas-

¹ The term "average" is frequently used as a general term to indicate any measure of central tendency. In its more restricted sense, it stands for "arithmetic mean" and will be so used here.

² The sum is usually designated by Σ (sigma), and the number of cases by N .

ures are grouped in a frequency distribution;³ and in that event the average is calculated by a somewhat different method.

The simple series.—The table below—problem I—shows a group of scores arranged in a frequency table, with each score set down and beside it the number of individuals who made that score (*f* column). It will be seen that one pupil scored 32, two scored 33, four scored 37, etc. To find the arithmetic mean of a frequency distribution of this type, it is necessary only to multiply each score by its frequency (number of times it occurs) and divide by the total number of cases. In the present problem

PROBLEM I

<i>Scores of a Class in History</i>	<i>Number of Cases (f)</i>
32	1
33	2
38	1
41	1
55	2
64	1
72	3
77	4
78	6
82	8
86	2
90	3
92	2
94	1
96	1
<i>Number of Cases</i>	<i>38</i>

³ In a frequency distribution the data are in orderly form, consisting of a series of *classes* of the measure and the number of items, or frequencies, in each class. See problem II, this chapter.

we have $(32 \times 1) + (32 \times 2) + \dots + (96 \times 1) = 2815$.
 $\frac{2815}{38} = 74.07$, which is the arithmetic mean.

The frequency distribution.—The grouping presented above, however, is practical only when the number of cases is not too great. But when as most frequently happens the number is large, even greater condensation is desirable. Instead of setting down each actual score, they may be grouped into class intervals. The interval may be 2, 5, 10, or any other number, depending upon a variety of considerations, such as the total number of cases and the range of scores (upper and lower limits). In the foregoing problem, if the scores were grouped with 5 as the class interval, we should have the following:

PROBLEM II

<i>Class Interval</i>	<i>Mid-point</i>	<i>Frequency</i>
30-34	32	3
35-39	37	1
40-44	42	1
45-49	47	0
50-54	52	0
55-59	57	2
60-64	62	1
65-69	67	0
70-74	72	3
75-79	77	10
80-84	82	8
85-89	87	2
90-94	92	6
95-99	97	1

Under an arrangement of this kind, the mid-point of each class interval is multiplied by the frequency of the interval, and the sum of the products is divided by

the number of cases. Thus $(32 \times 3) + (37 \times 1) \dots + (97 \times 1)$ divided by 38 = 74.1.

It will be noted, of course, that under this scheme the assumption is that the mid-point in each interval represents the scores in that interval.⁴ Thus, of the 10 cases scoring 75-79, it is assumed that 77 is the best single value for all 10. If this assumption is not to lead into error, it is necessary that the cases be symmetrically distributed in *each* interval. As a matter of fact, this is not always the case, but the errors throughout the distribution tend to counter-balance, so that the total error is really negligible. Comparing problems I and II, which deal with identical data we find that in I the sum is 2815, while in II, where the cases are grouped in class intervals, the sum is 2816. The respective averages are 74.07 and 74.1. In problems where the number of cases is very large, the saving in time and labor is very considerable when the method of class-intervals is used.

The short method.—The calculation of the arithmetic mean for a large number of cases may be further reduced in labor by employing what is known as the short method. Let us consider the data of problem II once more, simply as an illustration; for in practice the

⁴ What the mid-point in an interval is depends upon the nature of the units of measurement. In our problem above, the assumption is that the series is a "discrete" one; that is, 1 is the smallest unit, and there are no fractions. This would be the case if we were using "number of pupils," or "number of absences," or "number of failures" and the like. But frequently we have what is known as a "continuous" series; that is, a series where there are no gaps, as between 1 and 2, between 2 and 3, etc. We should have a continuous series if we were measuring height and weight, for example, and stating the results respectively in terms of inches and pounds with fractions. In the event of a continuous series, if we had a distribution in pounds as follows, 100, 105, 110, 115, etc., the midpoints would be 102.5, 107.5, 112.5, 117.5, etc.

Most of the teacher's calculations, it is safe to say, will deal with discrete series, as in our problems above.

short method would here not be necessary or desirable, inasmuch as the cases are so few in number. First, a guess is made as to the location of the mean; for purposes of computation, it is assumed to be at the mid-point of one of the intervals. It is not necessary that the interval containing the assumed mean be the one in which the real mean actually lies, although if it does happen so, the numbers involved in the calculations are smaller than otherwise. In our example let us assume that the mean falls at the mid-point of the 75-79 interval, that is at 77, which is then the "assumed mean" (AM). The third column, d , contains the differences, or deviations, from the assumed mean; that is, the number of units (whatever the scoring unit may be) that the mid-point in every interval is removed from the mid-point in the interval where the assumed mean lies. Thus the cases in the 70-74 interval are assumed to be

PROBLEM III

	<i>f</i>	<i>d</i>	<i>fd</i>
30-34	3	-45	-135
35-39	1	-40	-40
40-44	1	-35	-35
45-49	0	-30	0
50-54	0	-25	0
55-59	2	-20	-40
60-64	1	-15	-15
65-69	0	-10	0
70-74	3	-5	-15 (-280)
75-79	10	0	
80-84	8	+5	+40
85-89	2	+10	20
90-94	6	+15	90
95-99	1	+20	20 (+170)
$N = 38$		$\underline{38} - 110$	
		$c = -2.89$	

at 72, or 5 points below 77. Those in the 80-84 interval are placed at 82, or 5 points above, etc. Each value in the d column is then multiplied by its corresponding frequency, the products being written in the column headed fd . Each value under fd is, therefore, the product of the deviation (d) of each interval multiplied by the number of cases (f) having that deviation. The two totals are +170 and -280; the algebraic sum being -110. The algebraic sum is then divided by the number of cases (N), yielding the correction (c). The correction is then added or subtracted from the assumed mean, as the case may be. In this example $c = -2.89$; the mean (or average) is therefore 74.11 ($77 - 2.89$).

This method results in a distinct saving of labor where the frequencies are large. In simple problems, having relatively few cases, the first or second method herein described can be employed.⁵

THE MEDIAN.—The second measure of central tendency is the median, also very frequently found in educational and psychological literature. The median (abbreviated *Md.*) is that point on the scale of distribution on either side of which one half the measures fall. In a simple series where the scores are arranged in order of magnitude, the median is the *middle* measure.

The mid-score.—But best statistical practice dis-

⁵ There is yet another short method of finding the arithmetic mean. It is different only in one respect from the last method described. The values under the d column are divided by the size of the interval *before being set down*. In the above example the d column would read as follows beginning at the top: -9, -8, -7, -6, etc. When c is found it is multiplied by the size of the interval in order to get the value of the correction. Otherwise the methods are identical. This last method permits a saving in time and figuring when N is very large. Teachers, however, will in all likelihood find that the method described under problem III will meet their needs in practically all problems.

tinguishes between the two above definitions of the median; and rightly so. The latter is called the *mid-score* to distinguish it from the mathematically determined median. Thus, if there are 35 scores in a series, the eighteenth value is the mid-score; that is, it is the score on either side of which there are fifty per cent of the remaining scores. If the number of scores is *even*, then there is no one value that is the mid-score. In that event the average of the two mid-most scores is taken. Thus if we had the following percentages—65, 70, 71, 72, 76, 80, 85, 88—the mid-score would be the average of 72 and 76, or 74. It will be easily seen that if we were to add another score to the list, say 90, the mid-score would be 76. The mid-score is a convenient measure of central tendency when the number of cases is small and when the cases are not grouped in a frequency table.

Calculation of the median.—The median is always the point on the scale represented by $\frac{N}{2}$; that is, the number of cases divided by 2; and it will be the point on each side of which 50 per cent of the measures fall. The computation of the median will be illustrated in the following example. To obtain the median in this

PROBLEM IV

Score	f
95-99	2
90-94	5
85-89	8
80-84	9
75-79	6
70-74	5
65-69	3
60-64	2
N =	40

problem, it is necessary to count in 20 measures $\left[\frac{N}{2} \right]$ from either end. Starting from the lower end and adding, we have $2 + 3 + 5 + 6$, and it is found that there are 16 measures in the classes up to and including the 75 class. In other words, 16 scores are below 80. If the frequencies in the 80 class were added in, the number would be 25, which is too great. Clearly then, in order to obtain the median we must find the point in the 80 class below which 4 of the measures fall, for *up to* the 80 class we have only 16 cases; hence we must take the point of the fourth case in the 80 class as being the median point, then $16 + 4 = 20$, the desired median case. Since there are 9 measures in the 80 interval, and since they are assumed to be distributed at uniform steps in the interval, the desired point is $\frac{4}{9}$ of the way through the interval, or, in other words, $\frac{4}{9}$ of 5 units, which is $2\frac{2}{3}\%$ above its lower limit of 80. Adding $2\frac{2}{3}\%$ we find the median to be $82\frac{2}{3}\%$. The results obtained may be verified by counting from the other end, in this case from the upper, as follows: $2 + 5 + 8 = 15$. In order to locate the point above which the upper 20 scores lie, it is necessary to add 5 to the 15; that is, to come down $\frac{4}{9}$ of the way in the 80 interval. Thus $\frac{4}{9}$ of 5 are $2\frac{2}{3}\%$, which is the distance that we must go *down* into the 80 interval. Therefore, subtracting $2\frac{2}{3}\%$ from 85, we have the same result as before: $82\frac{2}{3}\%$ as the median.

We may summarize the steps for finding the median as follows: Divide the total number of cases by two $\left[\frac{N}{2} \right]$. Starting at either end of the distribution, the frequencies are added until the total is as large as possible with-

out exceeding $\frac{N}{2}$. The sum thus found is to be subtracted from $\frac{N}{2}$, and the result of this subtraction used as the numerator of a fraction. The denominator of the fraction is the number of cases (f) in the median interval; that is to say, in the interval immediately above or below, depending upon the direction in which the frequencies have been added, the last interval included in the addition. (In our example, the intervals which gave sums of 16 and 15, adding from low to high and then from high to low respectively.) The fraction is to be multiplied by the size of the interval (in our illustration it is 5). If the addition of frequencies has been from the lower end of the distribution, the product is added to the lower limit of the *median* interval; if the addition has been from the upper end down, the fraction is subtracted from the upper limit of the median class.⁶

⁶ FORMULAS FOR THE MEDIAN ARE AS FOLLOWS:

$$Md = l + \left(\frac{\frac{N}{2} - f_{up}}{f_{md}} \right) i \quad \begin{array}{l} \text{Median for distribution} \\ \text{counting up} \end{array}$$

$$Md = u - \left(\frac{\frac{N}{2} - f_{do}}{f_{md}} \right) i \quad \begin{array}{l} \text{Median for distribution} \\ \text{counting down} \end{array}$$

u and l = upper and lower limits of median interval; for example 80 and 85 in problem IV.
 f_{up} and f_{do} = total frequency up and down to median interval; e.g. 16 and 15.

f_{md} = frequency of the median interval; e.g. 9.
 i = width of interval.

Although the formula may be helpful to some, it seems most desirable to follow the steps as outlined above, rather than to adhere to a formula without a clear understanding of the members involved. After several practice problems, the derivation of a median will be found to be quite simple.

Advantages of mean and median.—The median is frequently taken as the “norm” in standardized tests. It is essential, therefore, that the teacher find the same central tendency where comparisons are to be made. Where the arithmetic mean is used as the norm, it should be found by the teacher in her own studies.

The median has certain other advantages besides being used frequently as a test norm. It is easily and readily computed. Where the series is a simple one (ungrouped), the mid-score may be readily observed. Furthermore, the median is not disproportionately affected by extreme measures.

The arithmetic mean, however, is used most frequently in problems where detailed statistical analysis is desired, for it may be found “on the way” to the computation of the standard deviation, which, in turn, is necessary for computing coefficients of correlation, both of which will be discussed in this chapter.

THE MODE.—The mode of a frequency distribution is that point on the scale where there are more frequencies (measures) than at any other point. It may therefore be regarded as representing the typical value or measure in the distribution. The mode may, in other words, be regarded as that measure which occurs most frequently in a distribution. In problem IV the mode is the mid-point of the 80 class interval because this interval contains the largest number of cases, 9; in problem III it is the mid-point of the 75-79 interval. Of all measures of central tendency the mode is the least frequently used. It is useful chiefly to indicate in a rough way the point of concentration. It may be located by observation.

At times it may happen that a distribution will

possess two points of concentration. Such a distribution is called "bi-modal." When there are more than two, the distribution is known as "multi-modal." When such conditions exist the meaning may be that there are two or more distinct groups, or that the "random sampling" is not a good one.

The mode should be employed only when the most frequently recurring measure is desired, or when a more or less rough approximation to the central tendency is wanted.

Limitations of the central tendency.—Before leaving the matter of central tendencies, we wish once more to emphasize the fact that these measures are useful and important as indicating *group* tendencies or properties; but they do not give us insight into the nature of any specific individual within the group. It happens too frequently that an individual member of the group is identified with the group's general characteristics, whereas he may be a marked variant. The discussion of measures of deviation will help in making this clear.

Measures of Deviation.—Thus far we have dealt only with the measures of central tendency, or the typical measures of a series. But data cannot be explained by a simple and single measure such as the central tendency. A further problem arises as to how the data are arranged around the central tendency. Are the values closely grouped, or are they widely distributed? The indices which will answer this question are variously known as measures of deviation, dispersion, or variability. The several measures of variability are the range, the quartile deviation, the median deviation, the mean deviation and the standard deviation.

THE RANGE.—As the term indicates, the range is the distance on the scale from the lowest to the highest score. It is obtained, of course, merely by subtracting the former from the latter. For example, the range of the series 60, 68, 72, 76, 83 . . . 97, is 37. It is usually also well to state the limits, such as 60-97 in this instance. This measure of variability is not very useful, for it gives only the limits of the distribution, without throwing any light on the degree of concentration around the central tendency.

Certain difficulties are encountered in the determination of the range in a frequency distribution if the lowest and highest scores are not actually known. The teacher, however, will not be faced with this problem since she will have the necessary information at hand.

THE QUARTILE DEVIATION.—The quartile deviation (Q) is one-half the distance between the first and third quartiles. The first quartile (Q_1) is the point below which one-fourth of the cases fall, and above which three-fourths of the cases fall. Similarly the third quartile (Q_3) is that point below which three-fourths of the cases lie and above which one-fourth of the cases lie. It should be noted that the median point corresponds to the point of the second quartile. The median point is $\frac{N}{2}$; therefore the point of Q_1 will be $\frac{N}{4}$, and the point of Q_3 will be $\frac{3N}{4}$. Q_1 and Q_3 are found in exactly the same way as the median, except, of course, that $\frac{N}{4}$ or $\frac{3N}{4}$ is taken instead of $\frac{N}{2}$, as the case may be.

The formula for the quartile deviation is

$$Q = \frac{Q_3 - Q_1}{2}.$$

It will be seen that when Q_1 and Q_3 have been determined, the range of the middle 50 per cent of the cases has really been set. This range is divided by 2 to obtain Q . If the distribution is symmetrical, Q will include 25 per cent of the cases; but this is not precisely true when the distribution is asymmetrical. However, in distributions only moderately asymmetrical, the error is very slight, and Q may be used with profit, inasmuch as it is readily determined and is most frequently employed with the median.

In our problem IV (above) $Q_1\left(\frac{N}{4}\right)$ is 75; $Q_3\left(\frac{3N}{4}\right)$ is 88.12. Therefore

$$Q = \frac{88.12 - 75}{2} = 6.56.$$

If the distribution is symmetrical we should interpret as follows: If we go out 6.56 units in both directions from the median (in this problem 82.22) we shall have the limits of the middle 50 per cent of the cases. As previously stated this is only very slightly in error when the distribution is moderately asymmetrical.

In general, then, the quartile deviation is that distance on the scale which, when laid off in both directions from the median, will designate the points within which lie the middle 50 per cent of the cases in the distribution.

THE MEDIAN DEVIATION.—As its name suggests, the median deviation (Md. D.) is the median of the deviations taken from a measure of central tendency. It may be found simply by tabulating the deviations and calculating their median. Like the quartile deviation, the Md. D. places the limits of the middle 50 per cent of the cases; but it is used ordinarily with the arithmetic mean. When the distribution is symmetrical the Md. D. and Q coincide. The former, however, is much less frequently employed than the latter.⁷

THE MEAN DEVIATION.—Here again it is clear that the mean deviation (M. D.) is the arithmetic mean of the deviations from the central tendency of the distribution. The M. D. may be determined from the arithmetic mean of the distribution or from the median. The computation is very simple, consisting only in finding the central tendency, then taking the deviation of each measure from the central tendency, summing these deviations without regard to sign and dividing by the number of cases (N). At times this involves the handling of large and cumbersome numbers, if the central tendency is not a round number, and if the distribution is long. For use in such an event there is a formula which reduces the labor very considerably. But we shall not present the details of this method here, since the teacher will find that the quartile deviation and the standard deviation meet her needs in practically any situation. In a classroom situation the number of cases is small enough

⁷ In a later section we shall discuss the standard deviation. The Md. D. is equal to .6745 of the standard deviation. One may therefore be determined from the other.

to permit computing the M. D. as outlined above.⁸

The M. D., when laid off in both directions from the central tendency, will set the limits which include approximately 57.5 per cent of the cases in the distribution.

THE STANDARD DEVIATION.—This index is probably the most important of the measures of variability and the most frequently used in research because of its greater reliability. It is abbreviated as S. D. or σ (sigma). The S. D. may be defined as the square root of the mean of the squares of the deviations taken from the arithmetic mean of the distribution. When the S. D. is laid off on the scale in both directions from the arithmetic mean, it sets the limits within which fall 68.26 per cent of the cases in the distribution, or roughly about two-thirds.

The simple series.—In a simple series (ungrouped) the formula is

$$\text{S. D.} = \sqrt{\frac{\sum d^2}{N}}.$$

To employ this formula we must first find the mean, then find the deviation of each case from the mean, square the deviations, divide their sum by the number of cases, and take the square root. Problem V (see page 364) illustrates the method.

⁸ For an explanation of the formula see Holzinger, *Statistical Methods for Students in Education*, pp. 102 ff.

PROBLEM V

Score	<i>d</i>	<i>d</i> ²
28	+8	64
25	+5	25
21	+1	1
20	0	0
20	0	0
19	-1	1
17	-3	9
17	-3	9
17	-3	9
16	-4	16
<hr/>		<hr/>
10 200		10 134
M = 20		S. D. ² = 13.4
		S. D. = 3.66

The S. D. here is 3.66. Knowing that the mean is 20, we may say, therefore, that approximately 68.26 per cent of the cases in this distribution fall between 16.34 ($20 - 3.66$) and 23.66 ($20 + 3.66$).

Frequency distributions.—Inasmuch as most problems in education are concerned with large numbers of cases, the method described above can hardly be used in such problems, particularly when the numbers run into the hundreds or even thousands. Although the above method will suffice for most problems of the teacher, it can become decidedly cumbersome, especially when the mean is not a whole number and when the cases run much above 30. The procedure now to be described is used when the cases have been grouped into a frequency distribution for which the formula is:

$$\text{S. D.} = \left(\sqrt{\frac{\sum fd^2}{N} - c^2} \right)^{\frac{1}{2}} i.$$

Suppose we have the following data:

PROBLEM VI

<i>Scores</i>	<i>f</i>	<i>d</i>	<i>fd</i>	<i>fd</i> ²
135-139 . . .	1 . . .	6 . . .	6 . . .	36
130-134 . . .	2 . . .	5 . . .	10 . . .	50
125-129 . . .	7 . . .	4 . . .	28 . . .	112
120-124 . . .	9 . . .	3 . . .	27 . . .	81
115-119 . . .	18 . . .	2 . . .	36 . . .	72
110-114 . . .	20 . . .	+1 . . .	20 (+127)	20
105-109 . . .	21 . . .	0 . . .	0 . . .	0
100-104 . . .	13 . . .	-1 . . .	-13 . . .	13
95-99 . . .	8 . . .	2 . . .	16 . . .	32
90-94 . . .	3 . . .	3 . . .	9 . . .	27
85-89 . . .	1 . . .	4 . . .	4 . . .	16
80-84 . . .	1 . . .	5 . . .	5 . . .	25
75-79 . . .	1 . . .	6 . . .	6 (-53)	36
<hr/>				
$N = 105$		$105 +74$		$105 520$
$c = + .70$		4.95 int.		
$c^2 = + .49$		$-.49$		
<hr/>				
$c = .70 \text{ int.}$		$\text{S. D.}^2 = 4.46 \text{ int.}$		
5		$\text{S. D.} = 2.11 \text{ int.}$		
$c = 3.5 \text{ units}$		$\text{S. D.} = 10.55 \text{ units}$		

It will be noted that the procedure is the same as that used in finding the mean by the short method, in which the deviations (*d* column) are divided by the size of the

⁹ The formula can be easily applied if it is noted that for every one of its members there is a distinct step in the process. The only new symbol here is the *i* which stands for the width of the class interval. In this example *i* = 5.

interval; that is, the deviations are stated in terms of *number of class-intervals* removed from the interval in which the mean is assumed to lie. With this in mind, then, it is clear that there is only one new column here which is not present in computing the mean: the fd^2 column. It is found, of course, by multiplying the corresponding values under d and fd . Thus, since there is one case which deviates six intervals, its fd value is 6 and its fd^2 value is 36. Two cases deviate five intervals, so that their fd value is 10, and their fd^2 value is 50, etc. The squares are summed up and divided by N . Now, to correct for the difference between the assumed mean and the true mean, c^2 is subtracted, giving S. D.²; then the square root is taken, resulting in the S. D., but *in terms of class-intervals*. If permitted to stand this way, the interpretation would be: "If we were to go 2.11 intervals in both directions from the mean, we should include 68.26 per cent of the cases." But we wish to have the S. D. in scale units; that is, in the units employed in making the measurements. Therefore, 2.11 is multiplied by 5 (number of units in each interval), and 10.55 is obtained as the S. D. of the distribution. In these data the assumed mean is 107 (the mid-point of the 105 interval); the correction is 3.5; the true mean is, therefore, 110.5. Our results thus indicate that 68.26 per cent of the cases fall within 110.5 ± 10.55 .

THE SIGNIFICANCE OF MEASURES OF DEVIATION.—In analyzing the data of any group, it is not enough to know the central tendency. The "spread" of the group must be known; that is, it should be known whether the group is relatively homogeneous or heterogeneous. If the de-

viation is large, then, of course, the group may be regarded as being composed of markedly unlike individuals. Consider, for example, a class-room in which the mean IQ is 105, but where the standard deviation is 20. This signifies that approximately two-thirds of the members of the class have IQs between 85 and 125! A considerable range. And it must be remembered that the remaining one-third are about equally distributed below 85 and above 125. This would be regarded as a heterogeneous class. By way of contrast consider another class having a mean IQ of 105, but with a standard deviation of 8 points. This class, because two-thirds range themselves between 97 and 113, may well be regarded as rather homogeneous, for the bulk of its pupils are of nearly like ability. As far as grouping is concerned it is decidedly superior to the first case cited. Yet, this discrepancy would not have been revealed without the S. D. In fact it is not unlikely that the correspondence of central tendencies might have misled some into regarding the groups as similar in composition.

THE COEFFICIENT OF VARIATION.—For the purpose of comparing the relative variability of two or more groups, there is what is known as the "Coefficient of variation," the formula for which is as follows:

$$\text{C. of V.} = 100 \frac{\text{S. D.}}{\text{M}}$$

This is the ratio of the S. D. to the mean; for convenience the quotient is multiplied by 100 in order to shift the decimal point. In the illustration above the results would be as follows:

$$(1) \text{ C. of V.} = 100 \frac{20}{105} = 19.04$$

$$(2) \text{ C. of V.} = 100 \frac{8}{105} = 7.61$$

The second group is, therefore, slightly less than 40 per cent as variable as the first $\left(\frac{7.61}{19.04}\right)$. The coefficient of variation is particularly valuable in comparing relative variability of groups when the arithmetic means are unlike, for then it is only by getting a ratio that a correct comparison can be made. Furthermore, this index may be used in situations where it is desired to compare the variability of the *same* group in different subjects. For example, the teacher or school officer may ask, "Are the pupils more uniform or less uniform in reading than in arithmetic?" Inasmuch as the averages in two different subjects are usually stated in different units, which are, therefore, not directly comparable, some comparing device is essential. The coefficient of variation serves this function, for the index of each, reading and arithmetic, may be found and directly compared.

A knowledge of the variability of intelligence and achievement within a class is extremely important for both teacher and supervisor. For the supervisor and the administrative officer, relative variability of different classes is important in reaching conclusions regarding a teacher's effectiveness and in formulating plans of classification.

USES OF MEASURES OF DEVIATION.—As has already been stated, the principal value of the measure of deviation for the teacher or school officer is the insight it gives into the homogeneity or heterogeneity of groups. But

measures of deviation are especially valuable in research problems of a practical and of a theoretical nature; problems which unfortunately can not be discussed in the brief compass of a single chapter. One such, of interest to the teacher, is the question of whether similar training tends to increase or decrease individual differences within a group. Another is the question whether pupils increase or decrease in variability when time for study is lengthened or shortened. In addition, the standard deviation is essential in the computation of other indices, such as the various coefficients of correlation.

Measures of deviation and coefficients of variation should be of particular interest and value to teachers and school officers in obtaining a better appreciation of the diversity of ability and actual achievement within a class or grade. These measures also offer a means for determining with scientific precision such matters as spread of chronological age within any group, length of school attendance, frequency of absences, spread of teachers' salaries, length of teachers' service in the system, range in size of classes, and a host of others. Not one of these could be dealt with at all adequately by means of central tendencies alone; for, as has already been shown, central tendencies are only a part of the story and may be decidedly misleading in the interpretation of data.

Correlation.—Perhaps no other single statistical device has found such widespread favor as the correlational procedure. Like most methods which gain very rapidly in popular favor, it has been used uncritically; and not infrequently broad unjustifiable conclusions

have been drawn. We have already emphasized the fact, in discussing central tendencies, that statistical procedures should be viewed chiefly as helpful instruments in the solution of a problem; but they must not be regarded as a panacea for all educational and psychological situations; nor can they be looked upon as anything but indices showing group characteristics, and, as in the case of central tendencies, not necessarily giving insight into the nature of any specific individual of the group. As we shall later see, the only possible exception to this statement may come when the correlation is perfect or "unity." For the teacher and the school administrator, the correlational technique has excellent possibilities for the purpose of yielding insight into the relationship between certain factors or qualities, as measured in groups. Such, for instance may be the study of the relationship between achievement in any two subjects; the degree of relationship between interest, personality traits, and the like, with school achievement. Likewise may be studied the correlation between height and weight, height and age, age and school grade, education of parent and intelligence of child; and many others. In fact any two sets of qualities, traits, or functions which can be measured may be correlated; and indeed that very thing has nearly been done. But it must be clear that the value of any coefficient of correlation depends upon the accuracy and validity of the measures employed and upon the soundness of the experimental procedure. In other words, here again it is necessary to go beyond the bare statistical index if our interpretations are to be as nearly valid as possible.

ROUGH CORRELATIONS.—As already indicated, the teacher may frequently desire to know such facts as whether in general the rapid pupil in arithmetic is also the most accurate. Is the rapid reader the one who also comprehends his reading best? Is the good student in algebra also a good student in geometry? Or, how does achievement in one subject compare with achievement in any other? There are several ways in which the desired information may be obtained. The pupils may be scored, let us say, on a test in arithmetic and on one in history. The pupils may then be divided in fourths in each subject test (or tenths, if the number is large enough), on the basis of the scores in each test. By inspection, then, it will be found that such and such a percentage of those in the highest fourth in arithmetic are also in the highest fourth in history; that a certain per cent in the highest fourth in arithmetic are in the second fourth in history, and so on for each division. In this way it is possible to form a general notion of the relationship between performance on the tests of arithmetic and history, and to get a general idea of the predictive value of success or failure in one subject for success or failure in the other.

THE SCATTER DIAGRAM.—A second method would be to construct what is known as a "scatter diagram." A table is prepared in which the intervals for one measure are shown horizontally, while the intervals for the second measure are shown vertically. Next, the scores or measures are arranged in pairs, and each pair is tabulated by placing a check mark in the proper square of the diagram. For example in the data below, pupil A scored 96 in an arithmetic test and 89 in an algebra test,

Pupil	Arithmetic Score	Algebra Score
A	96	89
B	99	96
C	93	91
D	86	79
E	73	78

etc. Taking these pairs of scores, and others, and marking them up in the correct squares, we have the following scatter diagram (Fig. 3).

		ALGEBRA											
		Score	50-4	55-9	60-4	65-9	70-4	75-9	80-4	85-9	90-4	95-100	Total
ARITHMETIC		95-100								/		/	2
90-4											/	/	2
85-9							/	/				/	3
80-4									/	/	/		3
C	75-9			/			/	/	/	/	/		7
	70-4				/		/	/					5
	65-9				/	/	/	/					6
	60-4		/		/	/	/						6
	55-9		/	/		/							5
50-4		/		/									3
Total		3	4	8	5	6	5	3	3	2	3	42	

Fig. 3

SCATTER DIAGRAM

Thus, pupil A scored 96 in arithmetic and 89 in algebra, so his check mark goes in row 95-100 and in column 85-89. The diagram may be read as follows: there is one pupil who scored between 95-100 in arithmetic and 85-89 in algebra; there is one who scored between 95-100 in arithmetic and 95-100 in algebra; there are two

who scored between 75-79 in arithmetic and 70-74 in algebra, and so on for every square. From the arrangement of the cases in the diagram the extent of the correlation between achievement in arithmetic and achievement in algebra may be estimated. If most of the cases are closely grouped about the diagonal from the lower left-hand corner to the upper right, the meaning is that individuals in general tend to maintain the same rank in one subject as in the other. The correlation is then said to be high and positive. The degree of correlation is estimated from the amount of scatter. If the check marks were scattered all about the diagram without any apparent order or arrangement, the correlation would be very low or zero. If the checks tended to arrange themselves closely along the other diagonal, from lower right to upper left, the correlation would be high and negative, or inverse.

THE COEFFICIENT OF CORRELATION.—But the inspection of diagrams can give only a very general idea of the relationship existing between the traits being examined. In order to achieve greater exactness and precision, a single index is used, *the coefficient of correlation*, abbreviated as r .

Let us suppose, for example, that we have measured 100 pupils for arithmetic ability and for ability in algebra. Suppose, further, that each pupil occupies the same relative position in each test; that is, the highest pupil in arithmetic is also highest in algebra, the second highest in arithmetic is also second highest in algebra, and so on down to the lowest pupil in arithmetic who is also lowest in algebra. Inasmuch as each pupil occupies the same relative rank in distribution, the

correlation is *perfect* and *positive*; that is, $r = +1.00$.

Consider now another group of 100 pupils who have been measured for intelligence with a group test, and for ability to discriminate weights. Suppose now, we find that a pupil with a high intelligence score is just as likely to get a low or mediocre score in weight discrimination as a pupil with a mediocre or low intelligence score; or that pupils with low or mediocre intelligence scores are as likely to get as high a rating in weight discrimination as those with high intelligence scores. In a situation like this, one measure is perfectly useless for purposes of predicting a score in the other. In other words, there is no apparent relationship between these two abilities, and the correlation is said to be zero; $r = 0$.

As a third illustration, consider a group which has been measured for speed in penmanship, and quality, or legibility. In this instance, for the purpose of illustration, we suppose the most rapid writer is the poorest in quality, the next fastest is second to the poorest, and so on until we come to the slowest writer who has the best score for quality. In this case we have a *perfect* but *negative* correlation, inasmuch as there is an exact inverse relationship between rate and quality of writing; $r = -1.00$.

Perfect and zero correlations, however, are not the rule; in fact they are rarely found in educational and psychological measurements. The correlations actually found range from $+1.00$, through zero, to -1.00 , in all degrees. The question naturally raised here is: "How large does r have to be to indicate a high degree of relationship, and how small must it be to indicate little or

no relationship? And what is the significance of the coefficients of various magnitudes?"¹⁰ Statisticians are not completely agreed as to the significance of various coefficients; yet, it seems reasonable to regard the following table as having merit.

Very high	± .90 to	± 1.00
High80 to	.89
Marked60 to	.79
Low40 to	.59
Doubtful20 to	.39
Insignificant	less than	.20

It is important to remember, however, that these divisions are arbitrary and that the dividing lines are not hard and fast. If two abilities show a correlation of .90, we may then say that the relationship between the two is very close; that is, the probabilities are very high that a pupil possessing a high degree of one will possess a high degree of the other; if he is mediocre in one, it is very likely he will be mediocre in the second; if he is poor in one, he will very probably be poor in the other; and so on all along the scale from excellence to failure. If the coefficient is about .80, the probabilities of a close relationship are still rather high; but when the coefficient falls to .70 and .60, though the relationship between the abilities is marked, prediction based upon these coefficients become decidedly less certain; for as the coefficient is more and more removed from unity ($r = 1.00$) its predictive

¹⁰ The significance of a coefficient of correlation is conditional upon various factors, such as the number of cases (N) and errors of sampling. These considerations we shall not enter into, for they are outside the province of this text. We are concerned primarily with a general interpretation for the teacher's purposes in arriving at a better understanding of class-room problems.

value decreases rapidly. If, for instance, it is found that $r = .60$ for arithmetic ability and ability in geography, it may be said that there is a marked *tendency* for pupils to maintain their relative ranks in the two subjects. Judging from this coefficient of .60, it is not very likely that an excellent pupil in arithmetic will fall to mediocrity, or lower, in geography, though it is not impossible; but it is not unlikely that he will fail to maintain the same degree of excellence in geography. Coefficients of correlation, in other words, indicate trends for groups. When r is fairly well removed from unity, the meaning is, then, that there are fluctuations in individual rankings in the two abilities or measures; it therefore is advisable in the study of problem cases to realize that individual pupils may be unique. Group indices, though extremely valuable, must be applied and interpreted with caution where the individual is concerned.

Not infrequently coefficients of correlation have been interpreted as showing cause and effect. There is nothing inherent in the correlation technique to justify such an interpretation. Any such relationship of cause and effect can merely be inferred as a consequence of the coefficient and of a knowledge of the materials which are being statistically treated. As an illustration, consider the studies of relationship between home environment (measured more or less objectively) and intelligence; a much discussed problem. The correlations found for these two factors are significant. Consequently, some have said that the superior home environment produces superior intelligence; others have maintained that the parents, being of superior intelli-

gence, endow their children with superior intelligence, and the superior environment is but a natural consequence of their intelligence. There is probably truth in both statements; but the r supports neither one nor the other. The coefficient of correlation indicates how far two sets of measures tend to vary together; or, inversely, when r is negative.

Computing the coefficient of correlation.—Computation of r is a rather complex process, though not too difficult. There are several different methods, but we shall confine our illustration to two of the simplest, since they will be adequate for ordinary class-room problems.¹¹

Suppose we have the following data:

PROBLEM VII

Pupil	Arith- metic	His- tory	Score(x)	Score(y)	dx	dy	d^2x	d^2y	xy	
A	95	90	+	25	+	16	625	196	+	400
B	90	94	+	20	+	20	400	400	+	400
C	85	86	+	15	+	12	225	144	+	180
D	80	78	+	10	+	4	100	16	+	40
E	75	66	+	5	—	8	25	64	—	40
F	70	74	—	0	—	0	0	0	—	0
G	65	82	—	5	+	8	25	64	—	40
H	60	62	—	10	—	12	100	144	+	120
I	55	70	—	15	—	4	225	16	+	60
J	50	54	—	20	—	20	400	400	+	400
K	45	58	—	25	—	16	625	196	+	400
Mean =	70	74					2750	1640	+	1920

$$r = \frac{1920}{\sqrt{2750} \cdot \sqrt{1640}} = +.90$$

¹¹ For a detailed treatment and explanation of methods for use with very large groups, any of the standard texts in statistics may be consulted.

STEPS IN THE COMPUTATION OF THE CORRELATION (PEARSON METHOD) OF TWO SETS OF SCORES FOR THE SAME GROUP OF PUPILS

First: Arrange the two sets of scores to be correlated in vertical columns, designated as x and y , with the two scores of each pupil opposite each other in the two columns. The scores in the x column should be arranged in ascending or descending order.

Second: Find the arithmetic mean for each column.

Third: Compute the deviation of each score for each column from the mean of that column, observing signs. These are to be designated in the dx and dy columns.

Fourth: Square each of the values in column dx and column dy to get the values for column d^2x and d^2y .

Fifth: Multiply each deviation under dx by the corresponding value of the dy column, observing signs. The algebraic sum of these products constitutes Σdxy , usually written $\Sigma x'y'$, which is the numerator of the fraction for the coefficient of correlation.

Sixth: Find the sums of column d^2x and d^2y , extract the square root of each sum and multiply one square root by the other. This gives $\sqrt{\Sigma x^2} \cdot \sqrt{\Sigma y^2}$, which is the denominator of the fraction yielding the coefficient of correlation.

Seventh: Divide the $\Sigma x'y'$ value obtained in *step fifth* by the value $\sqrt{\Sigma x^2} \cdot \sqrt{\Sigma y^2}$, obtained in *step sixth*. This gives the value of r .

It will be observed that there is really but one new column in this process, the $x'y'$ column. For the rest, it is as though the standard deviations were being found simultaneously for two sets of measures (see Problem V).

When an assumed mean is used, or when the mean is a fraction, making calculations with large decimals necessary, it is desirable to employ a method which

makes it necessary to introduce only one additional step; that is to adjust for the correction (c).¹² In this instance we should proceed exactly as above, but the deviations are taken from the *assumed mean*. It is then necessary to add column dx to obtain the correction for the x measure (arithmetic in this case), and column dy to obtain the correction for the y measure (history). The corrections for x and y are then incorporated into the following formula:

$$r = \frac{\frac{\Sigma x'y'}{N} - c_x c_y}{\sqrt{\frac{\Sigma x^2}{N} - c_x^2} \cdot \sqrt{\frac{\Sigma y^2}{N} - c_y^2}}$$

Reference to the discussion of the standard deviation will reveal that the two radicals in the denominator yield the S. D.; so that if one wishes he may calculate the S. Ds. and substitute them in the denominator.

This method, using the assumed mean, is recommended where the number of cases runs above 30 or 40, and where the actual arithmetic mean is a cumbersome mixed number.

The calculation of a coefficient or correlation may be mastered with a little practice; but for the teacher and the administrator, more important than the method of calculation is its interpretation. Sound interpretation comes primarily from a familiarity with the educational or psychological problems involved and from an appreciation of the two facts emphasized throughout this chapter: namely, the r indicates trends for groups, and it does not in itself indicate cause and effect.

¹² See discussions under arithmetic mean and standard deviation.

Graphic Method

It is generally advisable not only to arrange the results of tests in statistical form but also to present them in graphic form. Often facts and relationships not apparent in the statistical material are brought out in the graphs. Graphic methods present few special difficulties, and their widespread use makes it imperative that the class-room teacher be familiar with the construction and reading of curves.

The column diagram and the learning curve.—One of the simplest forms of graphic method used in education and business is here presented to show the percentage of children having playgrounds of various sizes. This graph is self-explanatory (Fig. 4).

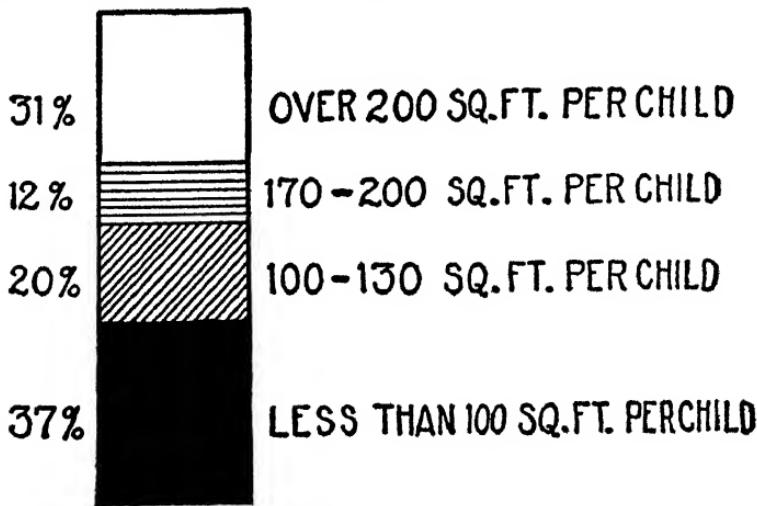


FIG. 4

SHOWING THE PERCENTAGE OF CHILDREN HAVING PLAYGROUNDS OF VARIOUS SIZES *

* From Statistical Methods Applied to Education, H. O. Rugg, p. 359.

Another common use of graphs is to show successive scores in some function. This is well illustrated by the well-known learning curve in telegraphy constructed by Bryan and Harter (Fig. 5). In this graph, time is represented along the horizontal axis, and the number of letters sent or received along the vertical axis. The

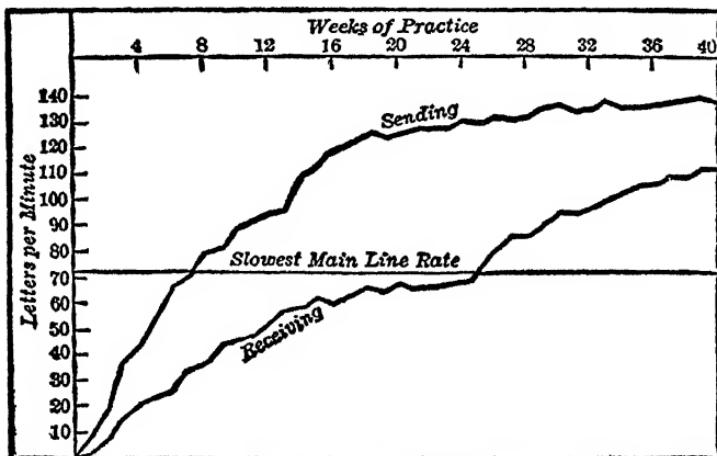


FIG. 5

CURVES OF LEARNING TO SEND AND RECEIVE TELEGRAPHIC MESSAGES (AFTER BRYAN AND HARTER)

rate for any amount of practice is shown by the point where the score line cuts the ordinate. For example, the sending rate for the twelfth week of practice is about 93 letters per minute and for the twenty-fourth week about 130 words per minute. In such curves, custom has established the zero point at the lower left hand corner. Time or amount of practice is generally represented along the horizontal axis and units of work along the vertical axis. A pictorial representation such as Figure

5 enables us to obtain a view of statistical facts which would not be immediately apparent from the data themselves; and it helps further in establishing a principle, if there be one. In this case the curve for sending indicates a rather rapid rate of improvement during the first weeks, the rate decreasing gradually, until after about the twentieth week there is very little improvement.

Curves such as those in Figure 5 are valuable also for purposes of comparison. For instance, if we wish to know whether there are sex differences in the learning of any type of material, we might plot a separate curve for each of the sexes and observe their characteristics. This method could be used for comparing groups in height, or weight, or any other measurable quality.

The following data may be used to illustrate the construction of a learning curve. A pupil practised silent reading 5 minutes per day for twenty days. His rate of reading in words read per second was as follows:

TABLE II

Day	Words per sec.						
1	2.34	6	2.94	11	3.40	16	3.68
2	2.00	7	3.00	12	3.44	17	3.80
3	2.58	8	3.28	13	3.52	18	3.95
4	2.75	9	3.32	14	3.60	19	3.80
5	2.90	10	3.00	15	3.68	20	3.92

These data, represented in Figure 6, show the characteristics usually found in learning curves. There is rapid progress at first, followed by less rapid gain during the latter part of the curve. At places in the graph there is practically no gain. These places are called

“plateaus.” Various reasons have been assigned for plateaus in learning. Swift¹³ and others attribute them to waning of interest or attention. Bryan and Harter¹⁴ explain plateaus as places where lower order habits are becoming automatized before more complex learning habits can be formed. After a certain amount of practice a point is reached where progress practically ceases. This is the permanent plateau and marks the limit of improvement. Whether the explanation of the plateaus be that of Swift or Bryan and Harter, the facts are clear.

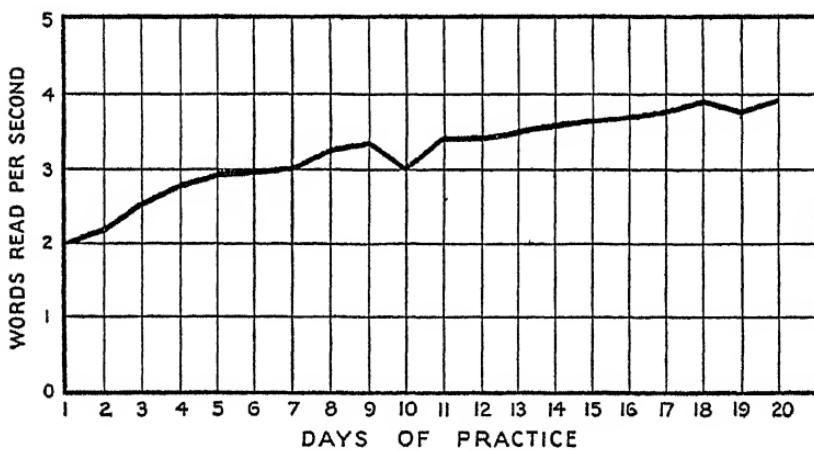


FIG. 6

GRAPH OF DATA PRESENTED IN TABLE II

This same type of curve may be used to represent the learning of a class as well as of an individual. In the case of a class, however, the levels of achievement (points on

¹³ Swift, E. J., *The Mind in the Making*.

¹⁴ Bryan and Harter, "Studies in the Physiology and Psychology of the Telegraphic Language," *Psychological Review*, Vol. 4: 27-53 and Vol. 6: 348-375, 1897-9.

the vertical axis) are determined by the central tendency of the group. This has the disadvantage of not representing the deviations from the central tendency. To overcome this disadvantage it is frequently desirable to draw three curves on the same chart: one curve for the median points, another for the points represented by the first quartile (Q_1), and still another for the third quartile (Q_3).

Frequency curves.—It has been demonstrated that measurements of natural phenomena as well as measurements of mental and social traits tend to distribute themselves symmetrically around the central tendency of the group. This is generally true unless some factors of selection have been operating to give a non-

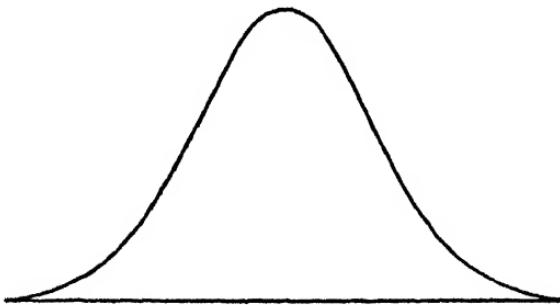


FIG. 7
NORMAL FREQUENCY CURVE

representative distribution. A symmetrical curve obtained from measurements of a non-selected group in mental, physical or social traits is the so-called *normal frequency curve* (see Fig. 7). Experiment has shown that such a curve, or a close approximation, will be obtained in psychological measurement (intelligence, edu-

cational achievement, rate of association, reaction time, etc.) in anthropometric statistics (height, weight, size of head), in the measurement of social and economic characteristics (wages, output, labor cost, birth rate, marriage rate), and in biological data (ratio of male to female births).

In problems of the class-room and the school it will be found that any set of measures, unless of a selected nature, will give a graph which approximates more or less closely the normal frequency curve. This will generally be true whether the measure be of pupils' height, weight, age, subject achievement, intelligence, etc., so long as the group is large enough and representative in nature. When statistics have been gathered for a group of pupils, it is desirable to present them graphically as well as by measures of central tendency and deviation. For this purpose there are two common types of frequency curves, either of which may be used. Consider, for example, the following data:

TABLE III

Scores of a class of 42 pupils on a test in algebra

<i>Scores</i>	<i>Frequency</i>
50-54	1
55-59	0
60-64	4
65-69	5
70-74	8
75-79	11
80-84	7
85-89	3
90-94	2
95-99	1

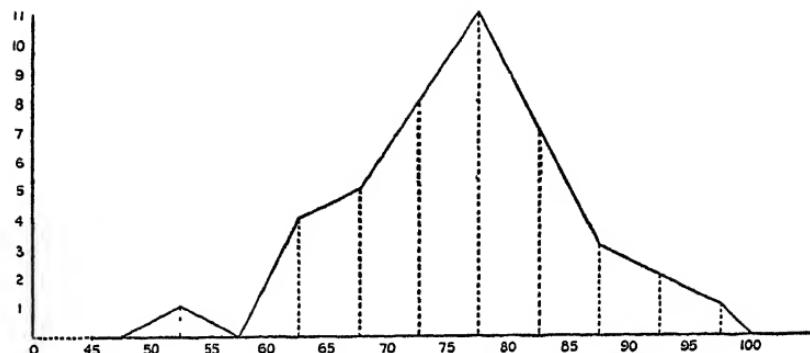


FIG. 8
FREQUENCY POLYGON

These data are represented by the curve in Figure 8 which is known as a *frequency polygon*.¹⁵ The same data may also be represented by the *histogram*¹⁶ in Figure

¹⁵ Steps in constructing a frequency polygon.

Draw two straight lines perpendicular to each other, the vertical line at the left of the graph paper, and the horizontal near the bottom, extending to the right. The vertical is known as the Y-axis, the horizontal as the X-axis. The points where the lines intersect is the origin.

Lay off the class-intervals of the distribution at regular intervals on the X-axis. As the origin, take the lower limit of the interval next below the lowest one in the distribution. Mark the successive points on the X-axis with the proper class-interval limits. Select as the unit on the X-axis a distance which will permit all the intervals to be represented on the one graph.

On the Y-axis lay off successive unit distances to represent the frequencies for the different class-intervals. These units should be so chosen as to permit the greatest frequency to be represented on the graph.

On the X-axis, from the mid-point of each class-interval go up in the direction of the Y-axis a distance equal to the number of frequencies of the step. Place a point there.

Join the points with straight lines. This will give the frequency polygon.

¹⁶ The histogram is a series of rectangles. Instead of placing a mark at the proper height above the mid-point of each class-interval, two perpendiculars are erected, one at each limit of the interval, their heights being equal to the frequency of the interval. The tops of each pair are connected by a straight line parallel to the X-axis. In practice the lines dividing the rectangles may be omitted and only the outline of the histogram drawn, as in Figure 8. Thus, two dots are placed where the upper corners of each rectangle should be, and the necessary lines are drawn to enclose the figure.

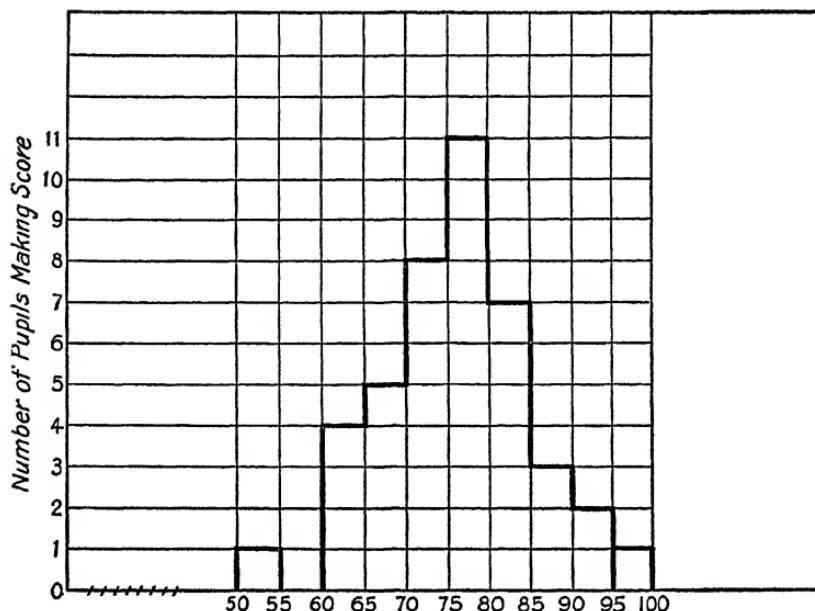


FIG. 9
HISTOGRAM

9. It is clear from these graphs that the scores on the algebra tests are rather well distributed and that the group is not a selected one.

SKEWED CURVES. It sometimes happens that a curve will show a concentration of cases at the upper or the lower end, because the group measured is not unselected and representative. In that event the curve is called *skewed* (Fig. 10). A skewed curve will result, for instance, from the measurements of the intelligence of a group of feeble-minded children, or from the measurements of a group of very superior pupils. We should likewise expect skewness if we were to measure arithmetic ability of eighth grade children by means

of a test designed for the sixth grade. In this case the scores would be piled up at the upper end of the scale, and the skewness would be *negative*. If, on the other hand, we were to examine the sixth grade pupils with a test designed primarily for eighth grade pupils, the scores would pile up at the lower end of the scale, and the skewness would therefore be *positive*. The presence of skewness in a curve indicates that some special condition exists with respect to the group. What that condition is can be discovered only through careful investigation into the situation; the curve itself does not offer the solution. A skewed curve found from the data of an achievement test may indicate unusually good or poor teaching, depending upon the direction of the skewness. It may indicate that the test is not a suitable one for the grade, being too easy or too difficult, as the case may be. Again it is possible that the class as a whole is well above or well below the average in general ability. In any event, when skewness appears, the meaning is that we are dealing with a selected and unusual situation, and that some factor or factors have been operating to produce that particular condition. What these factors might be can be determined only by analysis of the whole situation.

Through the use of graphs we are enabled to gain an insight into a distribution which might not otherwise be immediately apparent. In any case, however, they are an important supplement to frequency distributions, measures of central tendency, and deviations.

In this chapter we have considered only some of the elementary methods of statistics and graphic repre-

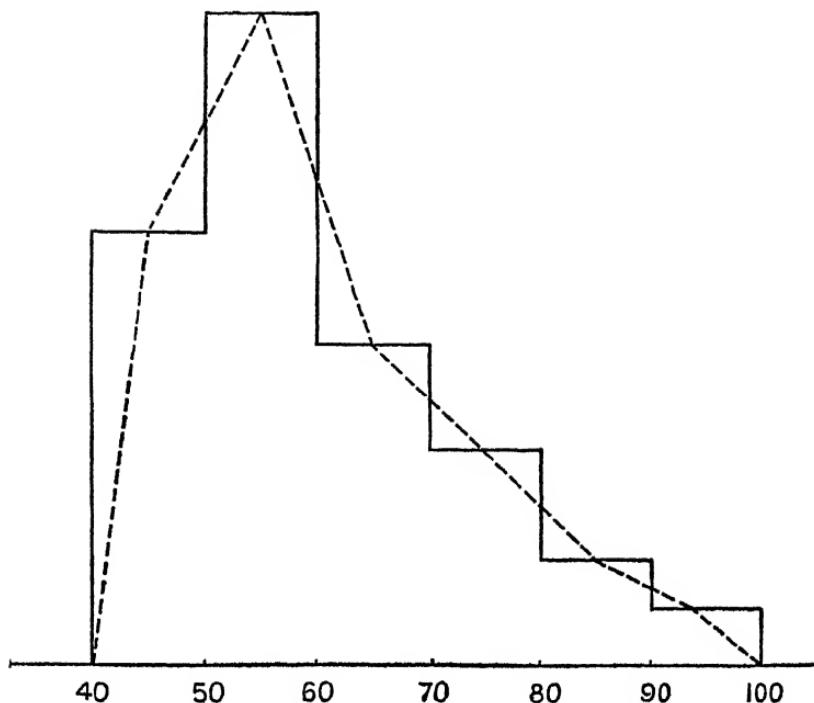


FIG. 10
A POSITIVELY SKEWED DISTRIBUTION

sentation. But the methods included should prove quite adequate for the teacher and administrator in analyzing and translating numerical facts into concrete and readily comprehensible form.

SELECTED REFERENCES

Chambers, G. G., *An Introduction to Statistical Method* (F. S. Crofts & Co., New York, 1925.)

Freeman, F. N., *Mental Tests* (Houghton Mifflin Company, Boston, 1926). Chapter III.

Garrett, H. E., *Statistics in Psychology and Education* (Longmans, Green & Co., New York, 1926).

Holzinger, K. J., *Statistical Methods for Students in Education* (Ginn and Company, Boston, 1928).

Odell, C. W., *Educational Statistics* (The Century Co., New York, 1925).

Otis, A. S., *Statistical Method in Educational Measurements* (World Book Company, Yonkers-on-Hudson, N. Y., 1925).

Rugg, H. O., *Statistical Methods Applied to Education* (Houghton Mifflin Company, Boston, 1917).

Thurstone, L. L., *Fundamentals of Statistics* (The Macmillan Company, New York, 1925).

INDEX

Abbott and Trabue, 175-177, 178
Accomplishment quotient, 62
Achievement Age, 59, 63
Achievement testing, 53
Achievement tests, practical uses of, 40-51; for diagnosis teaching, 42; evaluating methods of teaching, 44; diagnosis of classes, 45-47, diagnosis of individual pupils, 47-49; setting standards of achievement, 49, 50, promotion, 50, 51
Almack Tests in American History and Civics and Government, 243
American Council Tests in French, German, and Spanish, 304, 311
Civics and Government Tests, 243
European History Test, 243
French Grammar Test, 312
Solid Geometry Test, 277
Trigonometry Test, 277
Anderson, W. N., 96
Aptitude Tests, 342
Arithmetic 182-207; importance and problems of, 182, 183; types of tests, 183, 184; Courtis Standard Research Test, 184-187; Courtis Standard Practice Test, 187-189; Compass Diagnostic Test, 189, 190; Cleveland Survey Test, 191-194; Woody-McCall Mixed Fundamentals, 194-196; Monroe Diagnostic Tests, 196, 197; Monroe Standardized Tests, 198-200; New Stone Reasoning Test, 200, 201; comparison of Tests, 201-203; materials needed, 204; supplementary list of tests, 205, 206; selected references, 206, 207
Arithmetic mean, 349-354
Army Alpha and Beta, 36
Army Performance Scale, 341
Arthur Scale, 341
Ashbaugh, Ernest J., 89, 180
Average, *see* Arithmetic mean
Ayres, Leonard P., 60 n., 88, 96, 99, 120, 121
Ayres Measuring Scale for Handwriting, 108-111
Ayres Spelling Scale, 81-85
Badger Mechanical Drawing Test, 258
Bagley, W. C., 229 n.
Bagley, W. C., and Rugg, H. O., 229, 244
Bagster-Collins, 313
Ballou, F. W., 173, 178
Baltimore Plan, 49 n.
Barber, F. D., 280 n.
Barr Diagnostic Test in History, 225-228, 242
Barrett-Ryan Literature Test, 179
Batavia Plan, 49 n.
Beach Standardized Music Test, 258
Binet and Henri, 36
Binet and Simon, 36, 333
Binet-Simon Scale, revised, 333, 334
Blaisdell Instructional Test in Biology, 296

Bond, O. F., 313
 Branom's Diagnostic Tests in geography, 217-219
 Breed and Frostic Composition Scale, 173
 Brennan, Flora M., 250 n., 259
 Briggs English Form Test, 179
 Brinckley, S. G., 244
 Brooks, F. D., 155, 156
 Brown, A. W., 250 n., 259
 Brown and Coffman, 206
 Brown, M. D., and Haggerty, M. E., 180
 Brown-Woody Civics Test, 239-241, 242
 Brueckner, L. J., 310 n., 314
 Bryan and Harter, 383 n.
 Buckingham, B. R., 96, 205, 206
 Buckingham Extension of the Ayres Scale, 83-86
 Buckingham-Stevenson Place Geography Tests, 215, 216, 219
 Buckner, C. A., 18
 Burgess, May Ayres, 154, 156
 Burgess Reading Scale, 144-146
 Burt, Cyril, 39
 Burton Civics Test, 243
 Buswell, G. T., 125 n., 156, 205
 Buswell, G. T. and John L., 183 n., 206
 Buswell and Judd, 183 n., 184, 206

Camp, H. L., 297
 Carmichael, L., 127 n.
 Carter, R. E., 18
 Central tendency, variation in, 348, 349; limitations of, 359
 Chambers, G. G., 389
 Chapman-Cook Speed of Writing Test, 155
 Chapman, J. C., 65, 153
 Charters Language Test, 163-165, 178
 Chronological Age, 58
 Churchman, P. H., 314

Clapp, F. L., 183 n., 179, 205
 Clapp-Young Arithmetic Test, 205
 English Test, 179
 Clark Letter Writing Test, 179
 Clark-Ullman Test on Classical References and Allusions, 311
 Clem, O. M., 314
 Cleveland Survey Test, 184, 191-194
 Coefficient of Correlation, 373-379; computing, 377; Pearson method, 378, 379
 Coefficient of variation, 367
 Cole-Vincent Group Test of Intelligence for School Entrance, 344
 Coleman, A., 314
 Columbia Research Bureau Algebra Test, 277
 American History Test, 236-239, 242
 Chemistry Test, 296
 English Test, 179
 Physics Test, 291-293, 295
 Plane Geometry Test, 274-276
 Tests in French, German, and Spanish, 304, 305, 311
 Comin, R., 18
 Compass Diagnostic Tests in Arithmetic, 189, 190, 204
 Survey Tests, 205
 Contreras, Broom, and Kaulfers Test for Spanish Vocabulary, 313
 Silent Reading Test in Spanish, 313
 Cooprider Informal Exercises in Biology, 296
 Cornman, O. P., 96
 Cornog, J., and Stoddard, G. D., 207
 Correlation, 370-379
 Courtis, S. A., 48, 95, 118 n., 328
 Geography Tests, 216, 217, 219
 Music Tests, 252, 253
 Arithmetic Tests, 184-189, 204
 Tests in Handwriting, 116-120

Silent Reading Test, 146-148, 154
Cross English Test, 167, 168, 178
Current, W. F., and Ruch, G. M., 156
Curtis, F. D., 286 n., 297

Davis-Hicks Test in Roman History, 309, 311
in Historical Content and Background of Cæsar's Gallic Wars, 309, 311
Dearborn, W. F., 12 n., 18, 36, 65, 127 n., 156, 343, 344
Group Tests, 336, 337
Form Board Test, 341
Maze Test, 341
Defarri and Foran Test in Latin Comprehension, 313
Denny-Nelson American History Test, 243
Denver Curriculum Tests in American History and Government, 243
in Arithmetic, 205
in English, 179
in General Science, 296
in World History, 243
Detroit First Grade Intelligence Test, 36, 37
Mechanical Examination, 344
Reading Test, 155
Detroit (Engel) First Grade Intelligence Tests, 344
Deviation, 359-369; range, 360; quartile deviation, 360-361; Median, 362; Mean, 362, 363; standard deviation, 363-366
Dickinson, C. E., 156
Dickson, V. E., 65
Dougherty, M. L., 105
Douglas, H. R., 206, 276, 278, 282 n.
Algebra Tests, 267-280
Downing Science Tests, 282-284; 285
Downing, E. R., 297

Drawing, 246-248
Dvorak, A., 297
General Science Tests, 286-287, 295
Dykema, Peter, 259

Eastman School of Music, 251
Educational quotient, 62
Eells, N. C., 278
Examinations, inaccuracy of, 10-14

Fife, P. H., 314
Fillers, H. D., 180
Foran, T. G., 156
Diagnostic Computation Scales, 205
Forney Test in Map Reading, 220
Franzen, R. H., 65, 66, 180
Franzen, R. H., and Knight, F. B., 39
Freeman, F. N., 57 n., 62 n., 66, 101 n., 121, 206, 220, 389
Analytic Handwriting Scale, 103-107
Freeman, F. N., and Dougherty, M. L., 105 n.
Freeman, F. S., 30, 331 n., 345

Garrett, 28 n., 29 n., 39 n., 77 n., 390
Gary Plan, 49 n.
Gates Reading Test, Primary, 155
General Achievement Tests, 316, 317
Geography, 208-221; Hahn-Lackey Geography Scales, 209-211; Posey-Van Wagenen Geography Scales, 211-214; Buckingham-Stevenson Place Geography Test, 215, 216; Curtis Supervisory Tests in Geography, 216, 217; Brannon's Diagnostic Tests in Geography, 217, 218; materials needed, 219, 220; supplementary list of tests, 220; selected references, 220, 221

Gerry Test of High School Chemistry, 293, 295
 Gilliland, A. R., 103 n., 126 n.
 Glenn, E. R., 297
 Glenn-Osburn Instructional Tests in Physics, 294
 Glenn-Walton Instructional Tests in Chemistry, 293
 Godsey Latin Test, 307, 308
 Gordon, Hugh, 39
 Grading, 3-12, traditional methods of, 3, 4; need for more adequate method of, 4-10; teachers' need, 5; need in relation to the public, 5, need in relations with supervisory officers, 6; need in relation to pupils, 8; inadequacy of, 10-12
 Graphic Method, 380-389
 Graves Diagnostic Chart for Handwriting, 120
 Measuring Scale for Handwriting, 120
 Gray, C. T., 102 n., 120, 121, 156
 Gray Score Card for the Measurement of Handwriting, 115, 116
 Gray Silent Reading Test, 130-135
 Gray, W. S., 154, 156
 Gregory-Haggerty Geography Test, 220
 Gregory-Owens Test in Mediæval and Modern History, 243
 Gregory-Spencer Geography Test, 220
 Gregory Tests in American History, 243
 Grier Range of Information Test, 284
 Haggerty Intelligence Examination, 338, 343
 Haggerty, M. E., 36
 Haggerty Reading Examination, 149-152
 Hahn History Scale, 223-225, 242
 Hahn-Lackey Geography Scale, 209-211
 Handschin Modern Language Tests —French, 312
 Handwriting, 98-121; problems in, 98; quality, 98, 99, speed, 100-103; Freeman Analytical Scale, 103-108; Ayres Scale, 108-111; Thorndike Scale, 112-115, Gray Standard Score Card, 115-116; Courtis Standard Practice Tests, 116, 117; Methods of scoring, 117-120, materials needed, 120; supplementary list of tests, 120, 121; selected references, 121
 Hardy, R. E., 244
 Harlan Test of Information in American History, 228-230, 243
 Harris, E., and Breed, F. S., 278
 Hart, Diagnostic Tests and Drills in First Course Algebra, 277
 Geometry Test, 277
 Harvard Elementary Physics Test, 296
 Harvard Latin Test, 313
 Harvard-Newton Scale, 173
 Henmon, V. A. C., 299, 301 n., 303 n., 311
 Henmon Latin Test, 309
 Highsmith, J. A., 251, 259
 Hill Tests in Civic Information and Attitudes, 244
 Hill-Wilson Civic Action Test, 244
 Hillebrand Sight Singing Test, 258
 Hillegas Composition Scale, 173, 180
 Hillegas, Milo B., 171
 Hines, H. C., 66
 History, 222-245; problems of measurement in, 222, 223; Hahn History Scale, 223-225; Barr Diagnostic Tests in American History, 225-228; Harlan Test of Information in

American History, 228-230; Van Wagenen American History Scales, 231-233; Pressey-Richards Tests, 233-236; Columbia Research Bureau American History Test, 236-239; Brown-Woody Civics Test, 239-241; materials needed, 242-243; supplementary list of tests, 243, 244; selected references, 244, 245.

Hollingworth, L., 39, 96, 246 n.

Holzinger, K. J., 29 n., 39, 77 n., 363 n., 390

Horn, E., 79 n., 96

Hosic, J. F., 180

Hotz, H. G., 265 n., 277, 278

Hotz Algebra Scale, 264-266

Howell, W. B., 306

Hudelson, Earle, 96, 158 n., 178

Hudelson Composition Scale, 173

Huey, E. B., 126 n., 156

Hughes Physics Scale, 290, 291, 295

Hull, 345

Hunkins, R. V., and Breed, G. S., 207

Hutchinson Latin Grammar Test, 313

Hutchinson Music Test, 258

Illinois Algebra Test, 269-271, 277

Illinois General Intelligence Scale, 343

Indiana Composite Achievement Test, 328

Information Problem Tests in Geography, 220

Inglis, Alexander, 18, 181, 313

Intelligence quotient, 37, 58, 59, 63

Intelligence tests, 36-39, 330-345; nature of, 330-332; types, 332-335; representative group tests, 335-340; National Intelligence tests, 336; Dearborn Group Tests, 336, 337; Pintner-Cunningham Test, 337; Hag-

gerty Intelligence Examination, 338; Mentimeters, 338; Illinois Examination, 338; Mental-Educational Survey Test, 339; Otis Classification Test, 339; New Jersey Composite Test, 339; Terman Group Test, 339; Kuhlmann-Anderson Intelligence Tests, 339; Performance and Aptitudes Tests, 340-343; materials needed, 343, 344; supplementary list of tests, 344; selected references, 344, 345

Iowa Comprehension Test, 251

Iowa Physics Test, 280-288, 295

Iowa Placement Examinations, Revised Chemistry, 296

Iowa Placement Examinations, Revised, Mathematics, 277

Iowa Placement Examinations, Revised, Physics, 296

Iowa School Content Examination, 328

Iowa Spelling Scale, 89-90

Irmina, Sister M., 181

Johnson, F. W., 12 n., 18

Jones Spelling Scale, 90

Jordan, A. N., 314

Judd, C. H., 127 n., 156, 204

Judd, C. H., and Buswell, G. T., 156

Kansas City Scale of Handwriting, 120

Keener, E. E., 66

Kelley, F. J., 18

Kelley Mathematical Values, Test Alpha, 274

Kelley, T. L., Ruch, G. M., and Terman, L. M., 328

Kelley, Truman L., 60, 277, 278

Kennon Test of Literary Vocabulary, 180

Kepner, Background Test in Social Sciences, 244

Kinney Scales of Problems in Commercial Arithmetic, 205

Kirby (Iowa) Grammar Test, 180

Kirk, J. G., 121

Klapper, P., 181

Kline, L. W., and Carey, G. L., 256, 259

Knight, Luse, and Ruch, 207

Kohs, S. C., 345

Koos, L. V., 121, 244

Kwalwasser, J., 259

Kwalwasser-Ruch Test of Musical Accomplishment, 254-255

Lackey, E. E., 220

Language, 158-181; types of tests, 159; Starch Punctuation Scale, 160-161; Starch's English Grammar Scales, 161-163; Charter's Diagnostic Language Tests, 163-165; New York Survey Tests, 165-167; Cross English Test, 167, 168; Trabue Composition Scales, 168-170; Nassau County Supplement to Hillegas Scale, 171-173; Breed and Frostic Scales, 173; Hudelson Scale, 173; Harvard-Newton Scale, 173; Minnesota Composition Scale, 173; Seaton-Pressey Minimal Essentials Scale, 174; Pressey Diagnostic Test in English Composition, 174; Lewis English Composition Scale, 174; Abbott and Trabue Exercises in Judging Poetry, 175-177; materials needed, 179; supplementary list of tests, 179; selected references, 180, 181

Latin tests, 306-311

Latshaw, S., 245

Levine, A. J., and Marks, L., 345

Lewerenz Tests in Fundamental Abilities of Visual Art, 258

Lewis, E. E., 178

Lewis English Composition Scales, 174

Lippincott-Chapman Classroom Products Survey Tests, 328

Lister-Myer Handwriting Scales, 121

Logara-McCoy-Wright Tests for Appreciation of Literature, 180

Lohr-Latshaw Latin Form Tests, 313

Lord, E. E., 127 n.

Lunceford Diagnostic Test in Addition, 205

McCall, W. E., 60 n., 66, 142 n.

McMindes Plane Geometry Test, 277

Markham English Vocabulary Test for High School and College students, 180

Mathematics, secondary school, 261-279; problems of measurement in, 261, 262; Rogers Test of Mathematical Ability, 262-264; Kelley Mathematical Values, Test Alpha, 264; Ilotz Algebra Scale, 264-267; Douglas Diagnostic Test, 267-269; Illinois Algebra Tests, 269-271; Minnick Geometry Tests, 271-273; Schorling-Sanford Achievement Test, 273-274; Columbia Research Bureau Test, 274-276; materials needed, 276, 277; supplementary list of tests, 277, 278; selected references, 278

Mean, 28

Median, 28, 354-358

Meier-Seashore Art Judgment Test, 258

Mental Age, 57, 59, 63

Mental Tests, 35-38
Mentimeters, 338
Merrill-Palmer Tests, 341
Michigan Botany Test, 294, 295
Michigan Instructional Tests in Physics, 296
Miller Mental Ability Test, 344
Millikan, R. A., 280 n.
Minnesota Composition Scale, 173
Minnesota Mechanical Aptitude Test, 342
Minnick Geometry Tests, 271-273, 277
Minnick, J. H., 278
Mode, 358-359
Monroe Tests in Arithmetic, 196-200
Standardized Silent Reading Test, 139-141
Timed Spelling Test, 86-88
Monroe, W. S., 154, 184, 205
Monroe, W. S., and Buckingham, B. R., 328
Morrison, J. C., 273 n.
Morrison-McCall Spelling Scales, 93, 94
Multi-Mental Scale for Elementary Schools, 344
Multiple Skill First Grade Reading Scale, 155
Münsterberg Test, 342
Music, 246-260; as a special talent, 246-248; Seashore test, 248-251; Courtis Standard Supervisory Tests, 252-254; Kwalwasser-Krich Test, 254, 255; materials needed, 258; supplementary list of tests, 258; selected references, 259, 260
Myers, G. C., 207

Nassau County Supplement to Hillegas Scale, 171-173
National Committee on Mathematical Requirements, 261, 278

National Intelligence Tests, 336, 343
National Society for the Study of Education, 156, 245, 345
National Spelling Scale, 96
Nelson-Denny Reading Test, 155
Newcomb, R. S., 207
New Jersey Composite Test, 339
New Stanford Achievement Test, 317-323, 327
New Stone Reasoning Tests in Arithmetic, 200, 201
New York English Survey Tests, 165-167, 178
New York Latin Achievement Test, 313

Oakland Plan, 49 n.
O'Brien, J. A., 156
Odell, C. W., 66, 77 n., 390
Orleans Geometiy Prognosis Test, 277
Orleans-Solomon Latin Prognosis Test, 308, 309, 311
O'Rourke Mechanical Aptitude Test, 344
Osburn, W. J., 189 n.
Otis, A. S., 36, 390
Otis Classification Test, 326, 328, 339
Group Intelligence Test, 335, 343
Self-Administering Tests of Mental Ability, 344
Overlapping in Grades, 46

Peet-Dearborn Progress Tests in Arithmetic, 205
Performance Test, 340-341
Peters-Watkins Objective Tests for High School Physics, 296
Pintner-Cunningham Test, 337, 338
Pintner Mental-Educational Survey Test, 339
Pintner, R., 06
Pintner, R., and Marshall, H., "A

Combined Mental Educational Survey," 329

Pintner, R., and Patterson, D. G., 345

Pittsburgh Arithmetic Scales, 206

Porteus Maze Test, 341

Portland Plan, 49 n.

Posey-Van Wagenen Geography Scales, 211-214, 219, 221

Powers General Chemistry Test, 287, 288, 296

Powers General Science Test, 296

Powers, S. R., 297

Pressey, L. W., 121, 154, 174, 178, 258, 328, 344

First Grade Reading Test, 133-135

Test in Latin Syntax, 308, 311

Pressey-Richards Tests in American History, 233-236, 243

Public School Achievement Test in History (Orleans), 244

Punctuation Scales, 159-161

Pupils, factors in classification of, 64, 65

Rauth-Foran Chemistry Tests, 297

Reading, 123-157; importance and problems of, 123-127; types of tests, 127, 128; Gray test, 129-132; Oral reading check tests, 132, 133; Pressey tests, 133-135; Thorndike Visual Vocabulary Scale, 135, 136; Test of Word Knowledge, 137-139; Monroe reading test, 139-141; Thorndike-McCall Reading Scale, 141-143; Burgess Scale, 144-146; Courtis Silent Reading Test, 146-148; Gray Silent Reading Test, 148, 149; Haggerty Reading Examinations, 149-152; comparison of tests, 152, 153; materials needed, 154; supplementary list of tests, 155; selected references, 155-157

Reed, H. B., 96, 157, 221

Renfro's Diagnostic Tests in Plane Geometry, 278

Rice, J. M., 97

Rich Chemistry Test, 288, 296

Roback, A. A., 39

Rogers Test of Mathematical Ability, 262-264, 277

Ruch, G. M., 245

Ruch, G. M., and Crossman, L. H., 294, 296, 297

Ruch-Popenoe General Science Tests, 284-286, 296

Ruch and Stoddard, 28 n., 39

Rugg, H. O., 18, 229 n., 390

Rugg, H. O., and Clark, J. R., 266 n., 279

Russell-Harr Geography Tests, 220

Sammartino-Krause Standard French Test, 312

Sanford Tests, 342

Sanford, V. S., 279

Scale, definition of, 33

Schoen, M., 259

Schorling, R., and Clark, J. R., 279

Schorling-Clark-Lindell, 278

Schorling-Clark-Potter Arithmetic Test, 205

Schorling-Sanford Test in Plane Geometry, 273, 274, 277

Schutte English Diction Test, 180

Science, secondary school, 280-298; problems in the measurement of, 280, 281; Van Wagenen Scales, 281, 282; Dowling tests, 282-284; Grier Information Test, 284; Ruch-Popenoe test, 284-286; Dvorak test, 286, 287; Powers General Chemistry Test, 287, 288; Rich Chemistry Test, 288; Iowa

Physics Test, 288-290; Hughes Physics Scale, 290, 291; Columbia Research Bureau Physics Scale, 291-293; other tests, 293-294; materials needed, 295, 296; supplementary list of tests, 296, 297; selected references, 297

Seashore, C. E., 259, 260

Seashore Music Tests, 248-252

Seaton-Pressey Minimal Essentials Scale, 174, 178

Seattle Solid Geometry Test, 278

Seguin Form Board, 341

Shambaugh, C. G., and Shambaugh, O. L., 86, 97

Sims, V. M., 157

Sixteen Spelling Tests, 96

Social Studies in Secondary Education, 244

Sones-Harry High School Achievement Test, 323-326, 328

Spelling, 78-97; problems of, 79-81; Ayres Spelling Scales, 81-83; Buckingham Extension of, 83-85; Monroe tests, 86-88; Iowa Spelling Scale, 89, 90; Jones' study, 90, 91; Teachers' Word Book, 91, 92; Morrison-McCall Spelling Scale, 93, 94; materials needed, 94, 95; supplementary list of tests, 95, 96; selected references, 96, 97

Spink Grading Chart for Mechanical Drawing, 259

Standard deviation, 363-366

Standard tests, 20-39; factors involved in, 20-27; differentiation from traditional measures, 21, 22; necessary specifications, 22-30; criteria of a good test, 30-33; different kinds of measures, 33-35; mental tests, 35-38; selected references, 39

Stanford Achievement Reading Test, 155

Stanford Achievement Test—Arithmetic, 206

Stanford Revision of Binet-Simon Test, 36

Stanford Spanish Test, 313

Stanton, Hazel M., 250 n., 260

Starch, Daniel, 12 n., 13 n., 18, 97, 122, 181, 245, 246 n.

 Grammar Scales, 161, 162, 178

 Punctuation Scales, 178

Starch-Watters Latin Test, 313

Starch-Wise Scale for Measuring Handwriting, 121

Statistics, 346, 347

Stevenson-Coxe Latin Derivative Test, 313

Stevenson Latin Vocabulary Test, 313

Stevenson Problem Analysis (Arithmetic Reading Test), 206

Stoddard, G. D., 314

Stone, C. W., 184

Stone Narrative Reading Tests, 155

Stone Reasoning Test, 184

Supervisors' opinions, inadequacy of, 16, 17

Suzzallo, Henry, 97

Swift, E. J., 383 n.

Symposium, "Intelligence and Its Measurement," 345

Teachers' marks, unreliable, 14, 15

Teachers' opinions, inaccuracy of, 16, 17

Teachers' Word Book, 91-95

Terman Group Test, 339

Terman, L. M., 36 n., 39, 57 n., 58 n., 60 n., 80 n., 334 n.

Terry, P. W., 124

Tests, General Achievement, 316, 329

Thomson, G. H., 39, 345

Thorndike, E. L., 36, 37, 78 n., 86 n., 113 n., 120, 122, 139 n., 183, 207, 260, 270

Thorndike-McCall Reading Scale, 141-143, 154

Thorndike Scale for Handwriting, 112-115

Thorndike's Scale for the Merit of Drawing, 255, 256

Thorndike Visual Vocabulary Scales, 135-139

Thurstone, L. L., 390

Thurstone Vocational Guidance Test—Physics, 297

Tidyman, 97

Toops, H. A., and Symonds, T. W., 66

Torgenson-Fahnstock Music Test, 259

Trabue Completion Language Scales, 168-170, 178

Nassau County Supplement, 179, 181

Scales for Measuring Judgment in Orchestral Music, 260

Trabue, M. R., 66, 171

Trabue and Stockbridge, 344

Tressler English Minimum Essentials Test, 180

Tryon, R. M., 245

Twig French Vocabulary Test, 312

Tyler-Pressey Test in Latin, 308-312

Tyrrell American History Exercises, 244

Ullman-Kirby Latin Comprehension Test, 306, 312, 315

Vannest Diagnostic Test in Modern European History, 244

Van Wagener and Hubman and Patterson, German Reading Scale, 312

Van Wagener, M. J., 96, 173, 179, 180, 243, 244

General Science Scales, 281, 282

History Scales, 231-233

Variation, Coefficient of, 367

Wakefield Diagnostic Test, 180

Wallin, J. E. W., 97

Ward, C. F., 315

Weaver, C. T., 260

Webb Geometry Test, 278

West, A. F., 121, 315

Whipple, G. M., 35 n., 39, 155, 320, 336

White Latin Test, 307-312

Wildeman Standard Test in the Fundamental Operations with Common Fractions, 206

Wilkind Achievement Test in Spanish, 313

Prognosis Test in Modern Languages, 312

Williams, L. W., 279

Williams Primary Reading Test, 155

Willing, M. H., 179, 181

Wilson, G. M., 207

Wilson Language Errors Test, 180

Wilson Survey Test in Arithmetic, 206

Windes, E. E., and Greenleaf, W. J., 298

Winetka Plan, 49 n.

Wisconsin Inventory Tests in Arithmetic, 206

Witham Comprehensive History Tests, 244

Witham Standard Geography Tests, 220

Wood, B. D., 315

Woody, C., 217

Woody-McCall Arithmetic Test, 194-196, 204

World War, 36, 335, 336

Yerkes, R. M., 36 n., 336

Young, C. E., French and Spanish in the High School, 315

Young, Kimball, 345

Zaner, H. W., 121

